



# From Data to Decisions: Developing a Data Science Approach To Football Using Moneyball Principles.

By Kieran Henderson

Student ID: F015872

Loughborough University

23COP18: Data Science

Supervisor: Dr Chris Holland

Submitted on: 25th August 2024

## **Acknowledgements**

I would like to acknowledge Professor Chris Holland for the support that he has provided me and for answering all my questions. Chris has helped refine, plan and guide my ideas to create a data science project that I am proud of.

I would like to acknowledge my family for all their support and love throughout the process of my data science project.

I would also like to thank Jamie Graystone for his support in our final year of study.

# **Contents**

<b>Table of Figures . . . . .</b>	<b>4</b>
<b>Table of Tables . . . . .</b>	<b>9</b>
<b>Abstract . . . . .</b>	<b>8</b>
<b>Abbreviations and Acronyms . . . . .</b>	<b>9</b>
<b>Contents</b>	<b>13</b>
<b>Chapter 1. Introduction . . . . .</b>	<b>13</b>
<b>Chapter 2. Literature Review . . . . .</b>	<b>15</b>
<b>Chapter 3. Methodology and Design . . . . .</b>	<b>22</b>
<b>Chapter 4. Experiments, Evaluation, Results and Analysis . . . . .</b>	<b>27</b>
<b>Chapter 5. Discussion . . . . .</b>	<b>37</b>
<b>Chapter 6. Conclusions &amp; future work . . . . .</b>	<b>39</b>
<b>Reference List . . . . .</b>	<b>.41</b>
<b>Appendices . . . . .</b>	<b>56</b>

## Table of Figures

### Figures 3.1-3.4

$MAE = \frac{1}{N} \sum_{i=1}^N  y_i - \hat{y} $ <p>Where,  <math>\hat{y}</math> – predicted value of <math>y</math>  <math>\bar{y}</math> – mean value of <math>y</math></p>	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$
$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$	Normalized RMSE = RMSE / (max value – min value)

### **Figures 3.5-3.6**

#### Henderson Sports Analytic Data Model (Adapted from: Holland and Shaw (2019))

**Stage 1:** Data sources: sports analytics and data websites (Saherwardi, 2019, np.) (Holland and Shaw, 2019). The project's data sources will use sports analytics and data website FBref, which continually updates their data after each game played for on-field data (FBref, nd., np.) (André, 2023, np.). Moreover, a Kaggle dataset that has data web scrapped from Transfermarkt containing off-field data such as a player's transfer value will also be used (Transfermarkt, nd., np.) (Cariboo, 2024, np.).

**Stage 2:** Data Extraction: Web Scraping (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This project will webscrape the open source football data website FBref (nd., np.) and import the Kaggle transfermarkt dataset as a CSV (pd.read\_csv) (Cariboo, 2024, np.).

**Stage 3:** Data Preparation: Cleaning and Pre-processing (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This will include taking the FBref and Transfermarkt data and cleaning the data with processes such as removing duplicates, renaming columns into appropriate names, ensuring that the columns are the correct data types, etc (Cariboo, 2024, np.) (FBref, nd., np.).

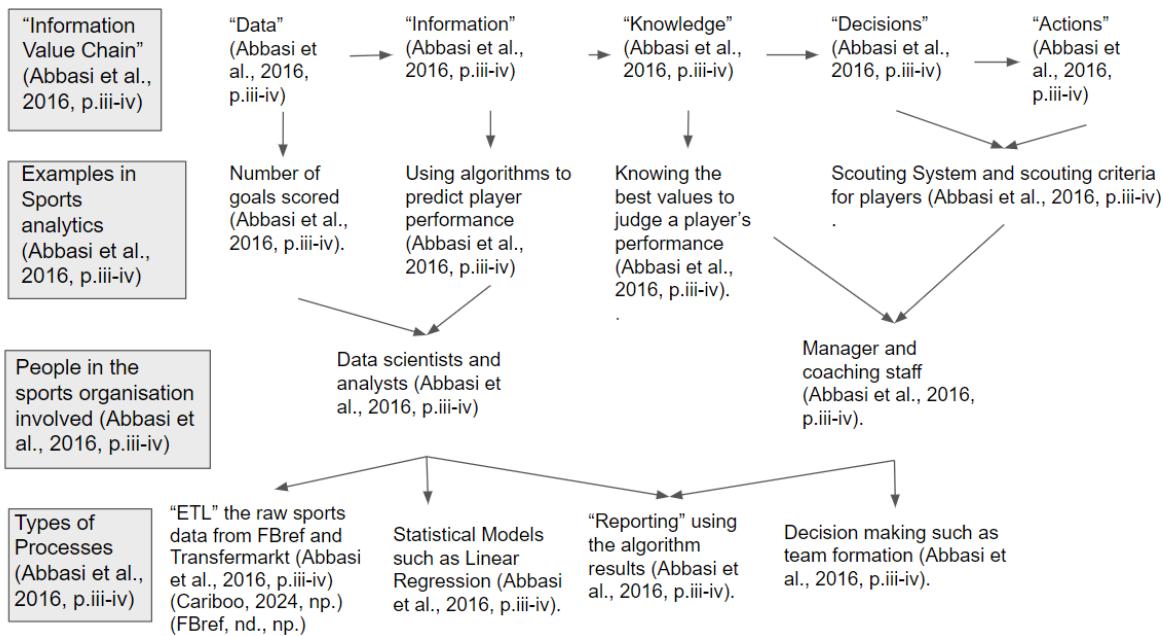
**Stage 4:** Data integration/building database (combining multiple data sources from different websites) (Saherwardi, 2019, np.) (Holland and Shaw, 2019). After preparing the data, the datasets can be “joined” together to make one big dataset with the relevant data. Integration will include taking “FBref” on-field data such as “goals” and combining them with “Transfermarkt” off-field data such as “estimated transfer value” (FBref, nd., np.) (Transfermarkt, nd., np.).

**Stage 5:** Data storage (structured data) (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This project will store the data as structured data in CSV files (Microsoft, nd., np.).

**Stage 6:** Statistical and Machine Learning Applications (Saherwardi, 2019, np.) (Holland and Shaw, 2019). In the author's project, an algorithm that will be used is Random Forest Regression in order to predict a player's performance using the processed and ranked variables that Random Forest used to predict a player's performance (Beheshti, 2022, np.). Metrics such as mean squared error will be used to evaluate the accuracy of the algorithm (Beheshti, 2022, np.).

**Stage 7:** Understanding Results and Outcomes (Saherwardi, 2019, np.) (Holland and Shaw, 2019). Using the data, we can analyse, evaluate, and discuss the results and outcomes, along with the aid of visualisation tools such as Tableau (nd., np.). For example, the author could compare the rankings of variables that random forest used to predict player performances and how similar they are to their own Player Score, a composite statistic used as a general means to evaluate a player's overall performance relative to their position.

### Sports organisation “Information Value Chain” (Adapted from: Abbasi et al., 2016, p.iii-iv)



### Figures 4.1.1-4.1.3

```

# Outfield Players' interface
Goals_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Performance_Gls:')
assists_slider = widgets.IntSlider(min=0, max=30, step=1, value=5, description='Performance_Ast:')
Tackles_slider = widgets.IntSlider(min=0, max=100, step=1, value=10, description='Tackles_Tkl:')
Age_slider = widgets.IntSlider(min=16, max=50, step=1, value=25, description='Age:')
npxG_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Expected_npxG:')
Progression_PrgC_slider = widgets.IntSlider(min=0, max=200, step=1, value=25, description='Progression_PrgC:')
Progression_PrgP_slider = widgets.IntSlider(min=0, max=200, step=1, value=25, description='Progression_PrgP:')
Usage_Rate_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Usage_Rate_Per:')
Score_slider = widgets.IntSlider(min=0, max=5000, step=1, value=500, description='Score:')
market_value_in_eur_slider = widgets.IntSlider(min=0, max=200000000, step=1000000, value=50000000, description='market_value_')
Tackles_TklW_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Tackles_TklW:')
Int_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Int:')
Blocks_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Blocks:')
Clr_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Clr:')
Challenges_Att_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Challenges_Att:')
Challenges_Tkl_per_slider = widgets.IntSlider(min=0, max=100, step=1, value=5, description='Challenges_Tkl_per:')
Challenges_Lost_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Challenges_Lost:')
Performance_CrdY_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Performance_CrdY:')
Performance_CrdR_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Performance_CrdR:')

# Goalkeepers' interface
Performance_Save_per_slider = widgets.IntSlider(min=0, max=100, step=1, value=0, description='Performance_Save_per:')
Penalties_saved_slider = widgets.IntSlider(min=0, max=20, step=1, value=0, description='Penalties_saved_Att:')
PSxG_minus_GA_slider = widgets.IntSlider(min=0, max=30, step=1, value=0, description='PSxG_minus_GA:')
Crosses_Stp_per_slider = widgets.IntSlider(min=0, max=30, step=1, value=0, description='Crosses_Stp_per:')
Number_of_Sweep_slider = widgets.IntSlider(min=0, max=10, step=1, value=0, description='Number_of_Sweep:')
Long_Pass_Acc_per_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Long_Pass_Acc_per:')


```

```

# Midfielders
# This function filters the Dataset based on the slider values made by the user
# This will use the sliders for the outfield players and apply them to the midfielders
def filter_players_mid(Goals, Assists, Tackles_Tkl, Age, Expected_npxG, Progression_PrgC, Progression_PrgP, Usage_Rate_Per, Score):
    scout_interface = (mid_df[(mid_df['Goals'] >= Goals) &
                               (mid_df['Assists'] >= Assists) &
                               (mid_df['Tackles_Tkl'] >= Tackles_Tkl) &
                               (mid_df['Age'] <= Age) &
                               (mid_df['Expected_npxG'] >= Expected_npxG) &
                               (mid_df['Progression_PrgC'] >= Progression_PrgC) &
                               (mid_df['Progression_PrgP'] >= Progression_PrgP) &
                               (mid_df['Usage_Rate_Per'] >= Usage_Rate_Per) &
                               (mid_df['Score'] >= Score)])
    )
display(scout_interface)

```

Goals:  5

Assists:  5

Tackles\_Tkl:  10

Age:  25

Expected\_npxG:  5

Progression\_PrgC:  25

Progression\_PrgP:  25

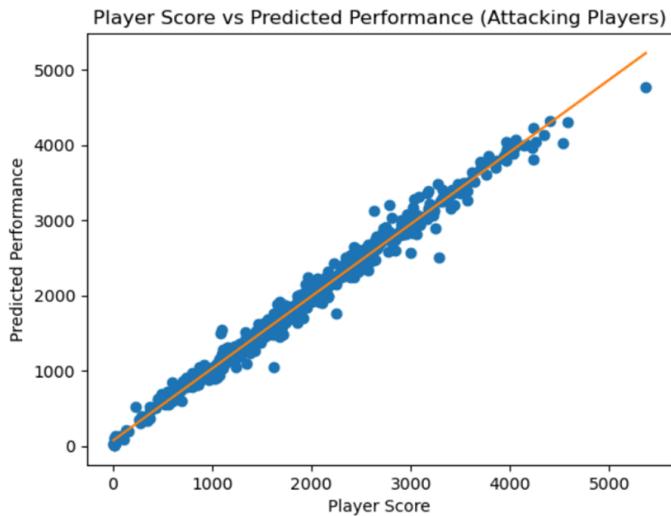
Usage\_Rate\_Per:  5

Score:  500

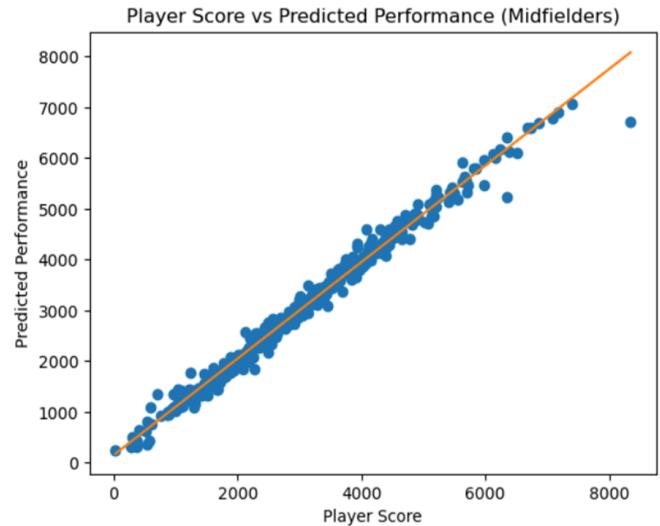
Squad	Player	Fouls_Committed	Fouls_Won	Offsides	Crosses	Penalties_Won	Penalties_Conceded	Own_Goals	Ball_Recoveries	... Height_in_C
3 Arsenal	Martin Ødegaard	30.0	36.0	3.0	77.0	1.0	0.0	0.0	158.0 ...	178
7 RB Leipzig	Xavi Simons	32.0	70.0	1.0	76.0	2.0	0.0	0.0	167.0 ...	179
12 Chelsea	Cole Palmer	23.0	26.0	6.0	81.0	1.0	0.0	0.0	105.0 ...	189
28 Nott'ham Forest	Morgan Gibbs-White	37.0	57.0	6.0	182.0	0.0	0.0	0.0	179.0 ...	171
30 Atalanta	Teun Koopmeiners	53.0	23.0	5.0	121.0	0.0	1.0	0.0	150.0 ...	184

## Figures 4.2.1-4.2.4

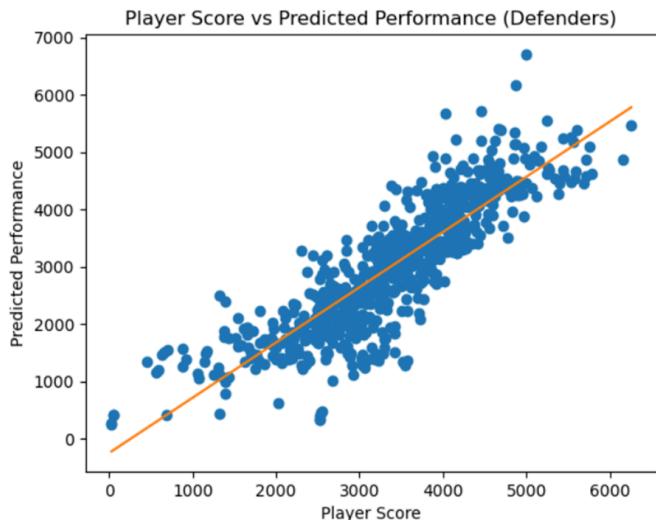
R-squared: 0.9868985587719865



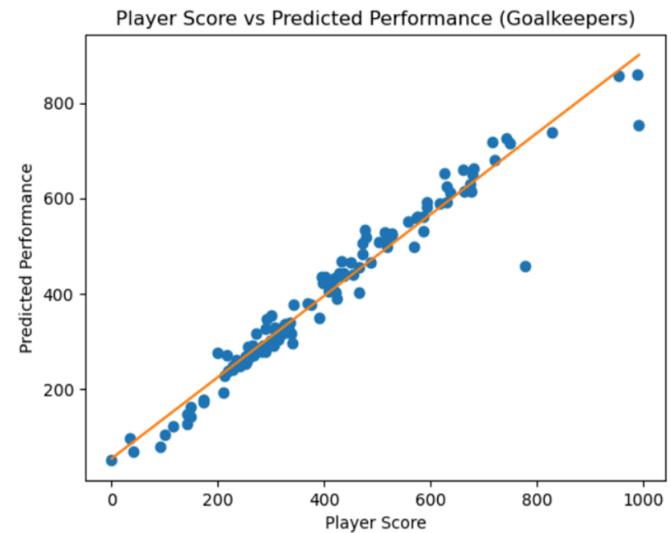
R-squared: 0.9865141285730494



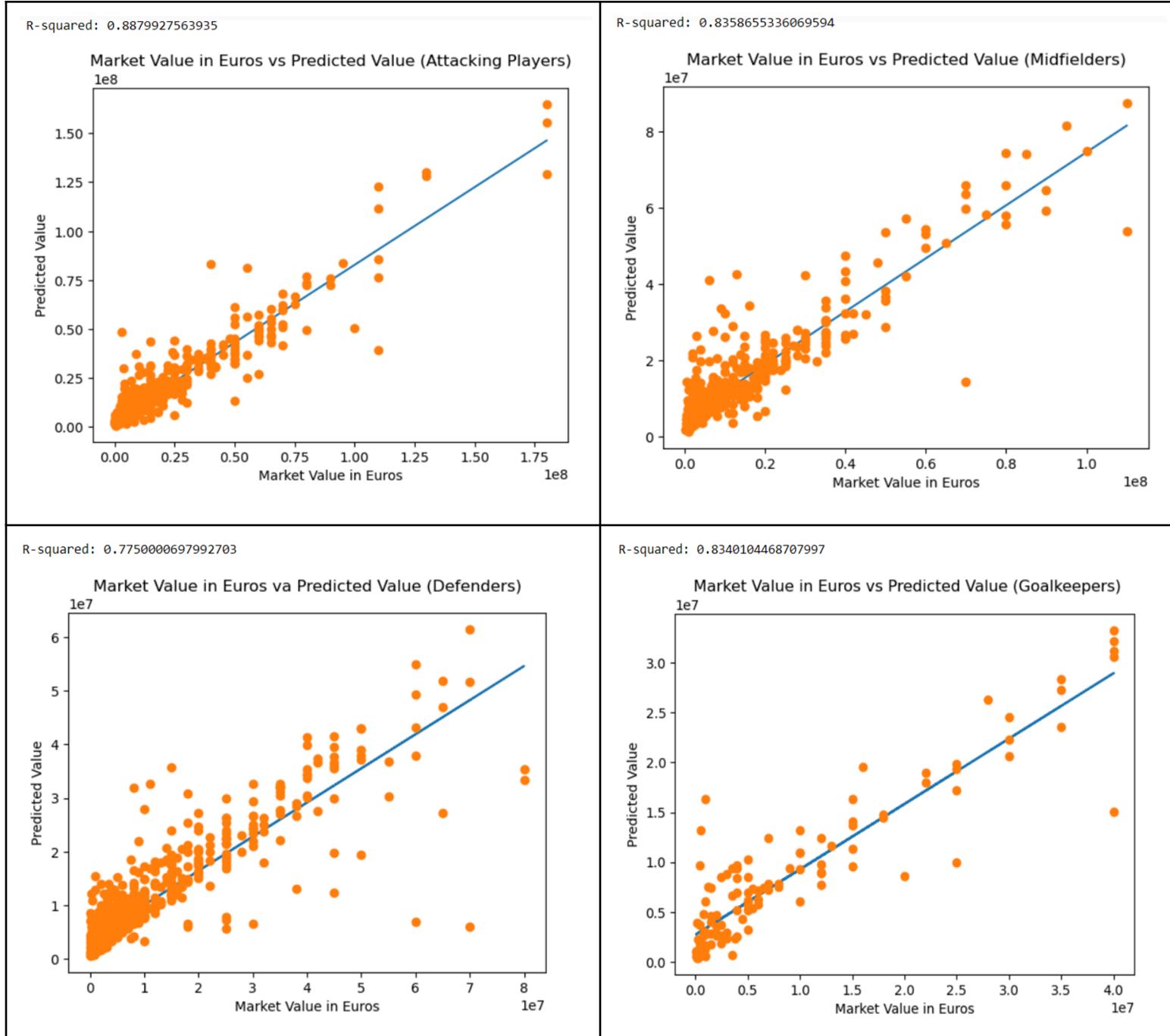
R-squared: 0.7420941519921884



R-squared: 0.9531838742623864



**Figure 4.2.5-4.2.9**



## Table of Tables

**Table 3.1**

	Attacking Player	Midfielders	Defenders	Goalkeeper
Number of rows	668	443	719	127
Number of columns	97	97	97	41

**Table 4.2.1**

Position	MSE	MAE	RMSE	Normalised RMSE
Attacking Players	4)882199	4)786	4)939	4)0.04
Midfielders	2)98545	2)221	2)314	1)0.04
Defenders	3)485808	3)539	3)697	3)0.03
Goalkeepers	1)7776	1) 54	1)88	2)0.09

**Table 4.2.2**

Position	MSE	MAE	RMSE	Normalised RMSE
Attacking Players	4)197790825073370	3)9309626	4)14063812	1)0.08
Midfielders	3)193532049707865	4)9960090	3)13911579	2)0.13
Defenders	2)170345637801076	2)7904020	2)13051652	3)0.16
Goalkeepers	1)67576926180288	1)5470278	1)8220518	4)0.21

**Table 4.2.3**

Position	MSE	MAE	RMSE	Normalised RMSE
Attacking Players	4)197790825073370	3)9309626	4)14063812	1)0.08
Midfielders	3)193532049707865	4)9960090	3)13911579	2)0.13
Defenders	2)170345637801076	2)7904020	2)13051652	3)0.16
Goalkeepers	1)67576926180288	1)5470278	1)8220518	4)0.21

## **Abstract**

As digital technology continues to evolve, sport institutions have begun to adopt this technology in order to gain a competitive advantage by using this technology along with new availability of big data to make decisions that rely on data driven means. Football is a sport that is beginning to embrace digital technology and big data, especially with the rise of tracking technology that can track a player's movement throughout an entire game and provide large volumes of data. Thus, the project will explore the literature review of sports analytics and the potential gaps that could be filled, then create useful managerial applications that could help contribute to the sport analytics research and literature, as well as discuss the benefits to managers, academia, etc. These applications will rely on machine learning and other data science techniques that can create apps that rely on data as input to provide outputs for management that could influence their decisions.

The project concludes that sport analytics can be used for managerial applications. This includes data sciences concepts, such as machine learning algorithms that can improve sport management's decision-making and give sports coaching staff that use these data driven and digital technology a competitive advantage compared to other teams, such as an improved player scouting.

## **Abbreviations and Acronyms**

**Artificial Intelligence - AI**

**Artificial Neural Networks - ANN**

**Data Analytics - DA**

**Decision Tree - DT**

**Expected Goal - xG**

**Machine Learning - ML**

**National Basketball Association - NBA**

**Predicted Performance - PP**

**Predicted Performance - PV**

**Player Efficiency Rating - PER**

**Player Score - PS**

**Random Forest - RF**

**Sports Analytics - SA**

**USP - Usage Rate Percentage**

**VORP - Value Over Replacement Player**

## **Chapter 1. Introduction**

The following Data Science Project focuses on sports analytics (SA) and the use of on-field and off-field analytics (Srivastava et al., 2021, p. 1-2). Data analytics (DA) and statistical methods have long been used in professional sports; however, with the publication of Michael Lewis' (2003) book *Moneyball*, there has become a new popularity and emphasis on statistics and DA in sports to gain a competitive advantage, especially with the use of advanced statistics and the availability of big data (Thakkar and Shah, 2021, p.825). However, as American sports continue to embrace analytics, football (or for Americans "soccer") has been slower to unleash the potential of SA, leading to questions such as: Can sport analytics programs be used in managerial applications to scout football players and differentiate types of players? What is an appropriate framework for measuring a player's performance, and can this be valuable for management? Can machine learning (ML) be used to predict the players' performances? Can "Moneyball" principles be used to find undervalued players in the football transfer market? (Lewis, 2003) (Szymanski, 2020, pp.57-59).

The research problem is essentially that football seems to be slower to adopt analytics compared to other sports, and this project aims to contribute to the literature by demonstrating the value of DA to sports management (Szymanski, 2020, pp. 57-59).

This project can be considered interdisciplinary, with disciplines of data science, sports management, DA, etc., all being used.

Peer-reviewed journals, blog posts, and business articles will be reviewed and critiqued on the use of SA, with the author's project contributing to the literature.

In addition to these literary sources, the project's datasets will be derived from the football data website "FBref," which contains tracking data of football teams and players, is open source, and will be collected through "web scraping," while a Kaggle dataset that has data that was scraped from "Transfermarkt" will also be used (FBref, nd., np.) (Transfermarkt, nd., np.) (André, 2023, np.) (Cariboo, 2024, np.).

As SA are used by management in sports organisations to gain a competitive advantage, the project will attempt to use "Moneyball" principles to gain insight and create managerial applications that can be used by sports organisations to improve their scouting, team selection, etc. (Lewis, 2003).

The managerial applications using SA will be:

1. A scouting system
2. An Estimated Market Value vs. Predicted Value and Player Score vs. Predicted Performance comparison graphs using the Random Forest (RF) Regressor algorithm (Beheshti, 2022, np.).
3. The Attack/Defence 4 quadrant grid

4. ‘Usage Rate Percentage’ and its’ impact on a player’s performance (Raumdata, 2020, np.).
5. Create a visualisation map of the nationalities with the highest quantity of talented players. Categorises the countries with the best talent into a “Moneyball Map” (Lewis, 2003).

This project will utilise a scouting system using football player data from multiple sources (Band, 2020, np.). The scouting system will be given inputs such as the desired player’s age, position, goals scored, etc., and the output will be a list of players that match their criteria and display their relevant data (Band, 2020, np.). Feature engineering, such as the generation of composite statistics and advanced metrics, will also be used to improve the scouting system’s potential (Hewitt and Karakuş, 2023) (Sarlis and Tjortjis, 2020).

The predicted value (PV) and predicted performance (PP) comparison graphs using the RF regressor will potentially demonstrate which variables and players have been overvalued or undervalued relative to their performance, which could allow management to find undervalued players that could improve their team without having to overspend their budget (Beheshti, 2022, np.).

These applications could be useful for management as they take data, which on its own might not be useful, and find ways to provide insight and allow management to make better decisions. The scouting system will allow a football scout to quickly create a list of players that meet a certain criterion through the use of data by giving an input of their desired metrics and an output of players. Finally, the use of ML can be used to predict the best performers and which metrics are the most important for predicting a player’s performance, which can be valuable to coaches who desire to find undervalued players and have a better idea of which variables can be used to find them. These applications could be valuable not only to sports management through better scouting and data-driven strategies, but also to academics in SA by demonstrating the value of analytics in sports outside the Americas. For players, this could be valuable as they could potentially be scouted regardless of their visual playing style, or rather, how well they play (Lewis, 2003, 2016, pp.13-49).

(Three additional managerial applications were created for this project in Appendix G.).

## Chapter 2

### 2.1 Literature Review Introduction

For this data science project, the research questions in the SA field are: Can sport analytics programs be used in managerial applications to scout football players and differentiate types of players? What is an appropriate framework for measuring a player's performance, and can this be valuable for management? Can ML be used to predict the players' performances? Can "Moneyball" principles be used to find undervalued players in the football transfer market? (Lewis, 2003).

In football, the "analytics movement" and field of SA can be viewed as underdeveloped due to multiple factors, such as football being a low-scoring game relative to other sports (Szymanski, 2020, p.58). Furthermore, with 11 players in a team, it is harder to measure an individual player's performance (Szymanski, 2020, p.58).

Therefore, football's struggle to keep up with other sports in the "data analytics movement" has left space in the literature for how SA can be used in football (Szymanski, 2020, pp.57-59).

Academic literature on SA has been growing, especially in American sports such as baseball and basketball, suggesting that they could provide valuable information to help answer the research questions in the context of football (Szymanski, 2020, p.58).

The literature review will start by exploring the field of SA by themes: first, the general concept of SA will be covered, including the use of AI and ML, followed by the application of SA in American sports, specifically basketball, a team sport that is more advanced in its development than football. This will then be followed by a review of the current literature for analytics in football. After that, ML methods will also be investigated, as will their application to both sports and non-sports analytical research. Finally, some off-field analytics will be examined.

## **2.2 The General Concept of Sports Analytics**

To start, Hintz's (2022, pp.1-7) paper examines the rise of professional sports management, embracing digital technology and data science methods. This starts with a summary of the rise of the “analytics movement/community” in the 1970s and 1980s, which was originally sports journalists and hobbyists who started to publish their data-driven work and then eventually posted it with the advent of the internet; eventually, it was also embraced by management in the 1990s and 2000s (Hintz, 2022, pp.1-7).

Moreover, Hintz (2022, pp.8-12) describes the process of a SA model, which is useful as it breaks down each process and could help guide one to creating a suitable SA model.

Beal et al.'s (2019) paper covers AI and ML techniques and how they have been applied to sports so far. These ML applications include tactics, player transfers, and predicting injuries (Beal et al., 2019). Then, Beal et al. (2019) assess the effectiveness of how AI and ML are used to improve teams' strategic and tactical decision-making, highlighting the benefits and downsides of each ML model and technique. Some ML methods used include neural networks, RF, DT, etc., with varying degrees of accuracy for each application (Beal et al., 2019, pp.12-20).

## **2.3 American Sports Analytics**

Now the literature review will move onto American SA, starting with Seshadri (2024, pp.1-3), who demonstrates an interesting concept by using ML techniques such as RF and ANN to improve advanced composite statistics such as basketball's advanced metric, Player Efficiency Rating (PER) (Hollinger, 2003, cited in: Seshadri, 2024, pp.1-3).

These ML models have unique capabilities to create a more accurate PER rating, which could help management make informed decisions backed by data. PER is a popular composite statistic used to assess a basketball player's overall performance; however, Seshadri (2024, pp.1-3) states that the metrics used in PER focus too much on offensive statistics such as points per game, while defensive statistics (blocks and steals) that can significantly impact a player's overall performance are less prioritised (Hollinger, 2003, cited in: Seshadri, 2024, pp.1-3). Seshadri's (2024, pp.1-3) research aims to improve PER by incorporating defensive contributions and by adjusting weightage through different ML models, thus making it a more effective statistic for judging a player's performance.

A potential limitation of the paper is that it is published in the Research Archive of Rising Scholars, which is not peer-reviewed, so despite it providing useful information and an interesting concept to build on, the paper's content has not been reviewed or scrutinised, suggesting some of the information is potentially not completely reliable or accurate (Research Archive of Rising Scholars, nd., np.) (Seshadri, 2024, pp.1-3).

Sarlis and Tjortjis' (2020, pp.1-2) paper explains and analyses the background of the advanced basketball statistics and other metrics used in the NBA and how different AI and ML can be

used. This is done to benchmark the current performance analytics used for evaluating player and team performance (Sarlis and Tjortjis, 2020, pp.1-2). This research uses advanced basketball statistics and ML algorithms to provide valuable information and insights for performance basketball analytics and to improve teams and players (Sarlis and Tjortjis, 2020, pp.1-2). Two supplementary articles about basketball metrics are Kubatko et al.'s (2007, p.1) paper, which defines the variables used in basketball statistics and gives explanations for each basketball metric and advanced statistics. This is then expanded upon by Terner and Franks (2021, pp.2-3), who discuss modern statistical and ML methods for basketball analytics and developments in the field since Kubatko et al.'s (2007, pp.1-3) publication.

The following literature explains why analytics can be considered superior to traditional scouting and non-statistical methods (Lewis, 2003). Michael Lewis (2016), in his book *The Undoing Project*, focusses on the work of Amos Tversky and Daniel Kahneman (psychologists whose work led to the rise of behavioural economics), as well as Daryl Morey, the Houston Rockets General Manager. This work can be supplemented by the articles by Thaler and Sunstein (2004), Tversky and Kanheman (1974, cited in: Berri and Eschker, 2005, pp.799-800), Collins (2018), and Packer (2018). Lewis' book details how NBA scouts have cognitive biases and fallible thinking using the ideas of Kahneman and Tversky and that simple statistical analysis can outperform "expert" judgement (Collins, 2018, np.) (Lewis, 2016, pp.13-49). For example, a human error due to bias could be "confirmation bias," as the Houston Rockets scouts encouraged Morey to reject Marc Gasol (a future NBA all-star) because he had "man boobs," i.e., he lacked an athletic build, despite Gasol being recommended by Morey's basketball recommendation model, demonstrating how humans' bias and subjective outlook can potentially lead to undesired outcomes and wasted resources, resulting in underperformance (Collins, 2018, np.) (Lewis, 2016, pp.13-49) (Packer, 2018, np.) (Thaler and Sunstein, 2004, pp.1390-1398) (Tversky and Kanheman, cited in: Berri and Eschker, 2005, pp.799-800). Therefore, the literature justifies the use of DA as it is both an efficient and effective method when coupled with traditional scouting that can make predictions that are not affected by cognitive bias or error (Collins, 2018, np.) (Lewis, 2016, pp.13-49) (Packer, 2018, np.) (Thaler and Sunstein, 2004, pp.1390-1398) (Tversky and Kanheman, cited in: Berri and Eschker, 2005, pp.799-800).

## **2.4 Football Analytics**

After analysing American SA, the literature review will move on to the current literature available in football and SA.

Hewitt and Karakuş' (2023, pp.1-3) research focusses on the advanced metric of Expected Goals (xG), as football is a low-scoring sport, making it harder to gather data compared to other sports, so additional parameters to examine the teams and players are important. xG is useful for tactics, recruitment, etc., because it measures how good a player is at getting into attacking positions or positions to score goals, potentially allowing scouts to observe a player's traits such as positional awareness or "football IQ" (Hewitt and Karakuş, 2023, pp.1-3).

Along with journal and business articles, blog posts have contributed to the field of SA (Hintz, 2022, pp.1-7). While not peer-reviewed, blog posts have the advantage of being quick, up-to-date, and less restricted when contributing to the field; for example, xG, a now widely used advanced football metric in academia and professional football punditry, was developed by bloggers (Hintz, 2022, pp.1-7) (Rathke, 2017, pp.514-515, p.525). Raumdata's (2020, np.) blog post adapts the basketball metric of Usage Rate Percentage ('USP'), a potentially useful metric, providing a formula on how to calculate it, allowing the author to conduct some feature engineering by using the provided formula to include a 'USP' metric.

Two football clubs that have recently embraced DA and the "Moneyball" approach are the English clubs, Brentford and Brighton, as detailed by Kuper (2024, np.), McDonnell and Sisneros (2023, np.), Bharadwaj (2023, np.), and Lemire (2023, np.) (Lewis, 2004). Both clubs are relatively small and have never won a trophy; however, using statistical methods, they went from low-value clubs with small amounts of resources to clubs cementing their status in the Premier League (the highest division in English football) (Lemire, 2023, np.) (Bharadwaj, 2023, np.). Both teams finished high relative to their resources, with Brighton even finishing above Premier League giants Tottenham and Chelsea, who spent far more, as Brighton finished 6th and achieved European football for the first time in their history (Kuper, 2024, np.) (McDonnell and Sisneros, 2023, np.) (Bharadwaj, 2023, np.) (Lemire, 2023, np.). Both team teams use statistical methods to acquire players in a "Moneyball" fashion, e.g., acquiring players with impressive advanced metrics such as xG that had gone undervalued in the market and would perform well in their teams' systems, leading to them increasing their value, which would sometimes result in the clubs selling them to richer clubs for a profit (e.g., Brighton sold Alexis MacAllister to Liverpool for £55 million, a £47 million net profit), which can then be reinvested back into the club (Kuper, 2024, np.) (McDonnell and Sisneros, 2023, np.) (Bharadwaj, 2023, np.) (Lemire, 2023, np.) (Lewis, 2003).

## **2.5 Random Forest and Other Potential Algorithms**

For the project, the primary algorithm that was used was the RF regressor (Kumar, 2020, np.) (Beheshti, 2022, np.). Kumar (2020, np.) and Beheshti's (2022, np.) articles explain and demonstrate the use of the RF algorithm and were useful in helping the author decide which algorithm to use, as RFs are a type of supervised learning algorithm that has the benefits of flexibility, simplicity, and accuracy as well as being able to handle big datasets. RF works by using multiple DTs, making them less prone to bias and overfitting (Kumar, 2020, np.) (Beheshti, 2022, np.). Furthermore, they are able to handle both categorical and numerical data, as well as non-linear data, making them ideal for this project (Kumar, 2020, np.) (Beheshti, 2022, np.). To examine their effectiveness, the author found Rodriguez-Galiano et al.'s (2015) article, which compared RF to other ML techniques such as neural networks, regression trees, and support vector machines in a study on “mineral prospectivity.” The paper concluded that RF was the most effective for the tasks provided, with the lowest MSE being the RF model (Rodriguez-Galiano et al., 2015, p.804, p.811, p.816). The paper also concluded that no generalisations can be made about the superiority of any ML method for different problems, as performance can differ by dataset, but in the paper’s application, RF was the most promising (Rodriguez-Galiano et al., 2015, p.804, p.811, p.816). Therefore, the author chose the RF method to use for the project due to its previously mentioned benefits and its effective use in disciplines not related to SA (Rodriguez-Galiano et al., 2015, p.804, p.816) (Kumar, 2020, np.) (Beheshti, 2022, np.).

## **2.6 Off-field Analytics**

The previously discussed literature mainly focused on on-field analytics, with references and limited detail on off-field analytics. This could be because SA mainly arose from a “grassroots” level as fans of the sports who wrote about the games they loved, who did not know about the inside dealings or the business aspects of the sport (Puerzer, 2005, pp.130-134) (Baumer and Zimbalist, 2014, p.ix) (Lewis, 2003). The sport only began to incorporate analytics after the popularity of the book *Moneyball* which was written by Lewis (2003) because he “loved the story,” as Billy Beane was the general manager of the Oakland Athletics baseball franchise, his role was strictly focused on the on-field issues of the organisation (Davenport, 2014, pp.10-13) (Puerzer, 2005, pp.130-134) (Baumer and Zimbalist, 2014, p.ix). Therefore, his use of analytics and their recognition in Moneyball (Lewis, 2003), led to their adoption by other franchises such as the Boston Red Sox, who won multiple titles (St. John, 2013, np.). This demonstrates that SA emerged at a “grassroots” level to focus on on-field analytics and that once it was adopted by sports organisations as a means to improve on-field performance, it eventually was also adopted for off-field issues (St. John, 2013, np.) (Davenport, 2014, pp.10-13) (Puerzer, 2005, pp.130-134) (Baumer and Zimbalist, 2014, p.ix).

Troilo et al. (2016, pp.72-4) explore the use of business analytics in professional sports. The first key finding of the paper is that using analytics by sports institutions does convince management that revenues are growing, which is backed by the next key finding, which is that these revenue increases occur and are significant, as sports institutions generally expect a 3% annual increase in revenue, and the research finding is that the adoption of analytics correlates with a revenue increase of 7.2% in the year after the application (Troilo et al., 2016, pp.81-82). However, a limitation of the paper, which the authors acknowledge as the limited time frame of the data, is that the results are derived from the effects after one year of adopting business analytics, so a longer time frame could strengthen the argument that analytics are the cause of the effects rather than just a correlation (Troilo et al., 2016, p.82).

Furthermore, another use of SA for off-field analytics is sports marketing (Holland, 2015, pp.1-2, pp.13-15). With the rise of digital technology such as social media, there has been a growing interest in using social media for sports marketing (Holland, 2015, pp.1-2, pp.13-15). Holland (2015, pp.1-2, pp.13-15) discusses how sports federations, leagues, and players have begun to embrace social media due to its growing audience and increasing revenue as an alternative to traditional broadcast rights. Sports stars have begun to market themselves more, and in some cases, such as Tiger Woods and LeBron James, they have outperformed their federations and leagues entirely, demonstrating the power of digital technology and data-driven decisions for off-field SA (Holland, 2015, p.13).

## **2.7 Literature Review Conclusion**

After reviewing the literature, one could observe that American sports such as baseball and basketball were much further ahead in the field of SA, with an emphasis on integrating digital technology and creating advanced/composite statistics that could help scout players, make team selections, improve tactics, etc. The review also found that there is a gap in the literature on DA in football (Beal et al., 2019, p.1) (Szymanski, 2020, p.58).

Basketball specifically is very ahead with its SA, as demonstrated specifically by Sarlis and Tjortjis' (2020), who's paper displays a huge glossary of advanced statistics and metrics used to measure a player's or team's performance and the embrace of ML to improve scouting, team selection, etc.

Football is behind in managerial applications such as the use of advanced and composite statistics, and very little research has been done on the use of ML to improve a football club's scouting, team selection, etc., especially with the availability of big data and datasets and the increasing use of new data sources such as tracking technology. Football's scepticism towards DA could be because, as one of the world's oldest competitive sports, it prides itself on tradition, with an early example being Arsène Wenger, who was an early proponent of sports science and the use of DA by using scientific methods and data to improve players performance, such as replacing traditional drills with training methods backed by sports science (Karen, 2016, np.). These were initially met with scepticism by players, managers, and the media, until Wenger's Arsenal kept winning trophies, leading to other teams phasing out their own traditional training methods with similar sports science-backed training drills (Karen, 2016, np.). This case study highlights how the sport has displayed a slow embrace of scientific and analytical methods in favour of "traditional" methods (Karen, 2016, np.).

Therefore, this project aims to contribute to the gaps in the literature by demonstrating the benefits of embracing SA, as well as provide some other managerial applications that could be used in football. The project also aims to expand the overall use of SA in football, specifically the underutilised tool of ML, as well as embrace the new available data sources (Beal et al., 2019, p.1) (Szymanski, 2020, p.58) (FBref, nd., np.) (Transfermarkt, nd., np.).

Bartosz et al. (2021, pp.4-8) use composite statistics; Hewitt and Karakuş (2023) use advanced statistics such as xG. So far, the author has been unable to find a paper that combines the two and aims to do this in this project to improve scouting, similar to how the PER metric works in basketball, in addition to using ML methods (Hollinger, 2003, cited in: Seshadri, 2024, p.1-2). These aspects, such as the analysis, managerial applications, methods, and algorithms that will be used for this project, will now be covered in the methodology chapter.

(See Appendix A for a table of the literature review sources, and Appendix B for supplementary material.).

## **Chapter 3. Methodology and Design**

Research Questions:

- Can sport analytics programs be used in managerial applications to find scout football players and differentiate types of players?
- What is an appropriate framework for measuring a player's performance, and can this be valuable for management?
- Can ML be used to predict the players' performances?
- Can “Moneyball” principles be used to find undervalued players in the football transfer market? (Lewis, 2003).

Data Sources:

"FBref" (nd., np.) is an open-source website that tracks the data of professional footballers. In this project, the data used will be men's footballers in Europe's top 5 leagues (Premier League, La Liga, Serie A, Bundesliga, and Ligue 1) for the 2023-2024 football season; therefore, it will be the player's most recent data at the time of writing (FBref, nd., np.). The datasets web scraped include “Standard Stats, Goalkeeping, Advanced Goalkeeping, Defensive Actions, Possession and Miscellaneous Stats” (FBref, nd., np.).

"Transfermarkt" (Transfermarkt, nd., np.), which contains tracking data of football teams and players, although this includes more off-field data such as their “estimated market value,” which is not included in the FBref data (André, 2023, np.). This dataset originated from Kaggle as a CSV file (Cariboo, 2024, np.). Unlike the FBref data, this includes players from other leagues; these excess players are removed during the ‘Final project code (Data Preparation).ipynb’ where all the rows with ‘na’ values are removed (FBref, nd., np.) (Cariboo, 2024, np.).

Theoretical framework, software and data:

Using the framework of SA and “Moneyball” principles, the project aims to find the best players statistically and according to quantitative data, so qualitative aspects such as the most visually appealing players will be ignored to avoid subjective bias (Lewis, 2003).

The three Moneyball principles are: 1) Using data analytics and statistical methods 2) Using these analytics and methods to find players with strong performances statistically, regardless of the aspects of their visual performance 3) Find the players who have strong performances but are undervalued, i.e., players with low transfer fees (Lewis, 2003).

The software that will be used is Python (through Jupyter notebooks) and Tableau due to their simplicity, user-friendliness, and versatility relative to other programming languages and visualisation software (Python, nd., np.) (Jupyter, nd., np.) (Van Deusen, nd., np.) (Tableau, nd., np.).

The data collection method used for this project will be web scraping, which will be used to extract various player statistics from the football data website "FBref," (nd., np.) which contains tracking data of football teams and players and is open source, while a Kaggle dataset that has data that was scraped from "Transfermarkt" (nd., np.) will be used (André, 2023, np.) (Cariboo, 2024, np.). Using this extracted data, a new dataset will be saved as a Microsoft Excel file (Microsoft, nd., np.). These two data sources, when joined together, create a large and detailed dataset with different variables to be used for analysis (Cariboo, 2024, np.) (FBref, nd., np.).

The project contains over 100 variables, all used in football, to determine the quality of a player, their "Player Score," their ideal role in the team, etc. Some key variables include "Goals," the metric that determines who will win the game by whoever "scores" the most and "Market Value in Euros," which is the estimated value a player would be sold for in the football transfer market as estimated and calculated by the website "Transfermarkt" (Transfermarkt, nd., np.).

(For the full list of variables, see Appendix C.).

Number of columns and rows for each processed dataset:

**Table 3.1**

	Attacking Player	Midfielders	Defenders	Goalkeeper
Number of rows	668	443	719	127
Number of columns	97	97	97	41

Another data analysis method is creating a new composite statistic that uses metrics such as goals scored, xG, tackles won, etc. (FBref, nd., np.). This new value, inspired by the basketball statistic "PER" and other advanced basketball statistics, can allow for significantly better identification of the best players (Basketball Reference, nd., np.). This will be named "Player Score," (PS) and there will be four types of PS, one for each position.

Basketball does not do this by position; PER is measured the same regardless of position (Basketball Reference, nd., np.). This project will split the football equivalent of "PER" into the 4 "main" positions to distinguish the nuance between the positions, with the variables being used and/or weighted differently for each position (e.g., tackles are more important and therefore weighted more highly for defenders compared to attackers) (Basketball Reference, nd., np.).

(The criteria for each position's PS is available in Appendix D.).

For making predictions and performing regression, the type of ML method used will be RF; e.g., when predicting the value of a football player, the data will be put in a RF regressor (Beheshti, 2022, np.). For the RF model to be evaluated, finding its accuracy is essential (Chugh, 2020, np.). The metrics that will be used to measure the accuracy will be Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and normalised RMSE (Chugh, 2020, np.) (Bobbitt, 2021, np.). These metrics can all be compared to one another to determine which position the RF model was the most accurate for, with an emphasis on the RMSE and normalised RMSE, as they are the most reliable and useful metrics (Chugh, 2020, np.) (Bobbitt, 2021, np.).

Linear regression will also be used, with the interpretations of the correlations being found in Schober et al. (2018, p.1765).

The equations for these metrics can be found in the figures below (Chugh, 2020, np.) (Bobbitt, 2021, np.):

### Figures 3.1-3.4

$MAE = \frac{1}{N} \sum_{i=1}^N  y_i - \hat{y} $ <p>Where,  <math>\hat{y}</math> – predicted value of <math>y</math>  <math>\bar{y}</math> – mean value of <math>y</math></p>	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$
$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$	$\text{Normalized RMSE} = \text{RMSE} / (\text{max value} - \text{min value})$

For the programming, these will be the hyperparameters and will be adjusted if needed:

RF hyperparameters: n\_estimators=100, max\_depth=30, min\_samples\_split=2, min\_samples\_leaf=1, random\_state=42 (Beheshti, 2022, np.).

A potential limitation of the dataset is that there are some important variables found in neither of the sources, such as “duels won.” While this is a limitation, it should not have a profound impact on the project as a whole, as similar metrics such as tackles, interceptions and aerial duels won are still included.

This project will be similar to Comscore as it collects a variety of online consumer behaviour such as websites visited, time spent on each website, etc., which is also referred to as “clickstream data” (Ferguson, 2014, cited in: Holland et al., 2020, p.57) (Wixom et al., 2013; cited in: Holland et al., 2020, p.57), with the scouting recommendation system working on a similar basis as it collects the players' data, e.g., their touches per game, tackles, etc. (Holland et al., 2020, p.57).

Furthermore, using the Holland and Shaw (2019) data model on Comscore, the author has adapted the data model for the project.

(For football terminology, see Appendix E.).

## Figures 3.5-3.6

### Henderson Sports Analytic Data Model (Adapted from: Holland and Shaw (2019))

**Stage 1:** Data sources: sports analytics and data websites (Saherwardi, 2019, np.) (Holland and Shaw, 2019). The project's data sources will use sports analytics and data website FBref, which continually updates their data after each game played for on-field data (FBref, nd., np.) (André, 2023, np.). Moreover, a Kaggle dataset that has data web scrapped from Transfermarkt containing off-field data such as a player's transfer value will also be used (Transfermarkt, nd., np.) (Cariboo, 2024, np.).

**Stage 2:** Data Extraction: Web Scraping (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This project will webscrape the open source football data website FBref (nd., np.) and import the Kaggle transfermarkt dataset as a CSV (pd.read\_csv) (Cariboo, 2024, np.).

**Stage 3:** Data Preparation: Cleaning and Pre-processing (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This will include taking the FBref and Transfermarkt data and cleaning the data with processes such as removing duplicates, renaming columns into appropriate names, ensuring that the columns are the correct data types, etc (Cariboo, 2024, np.) (FBref, nd., np.).

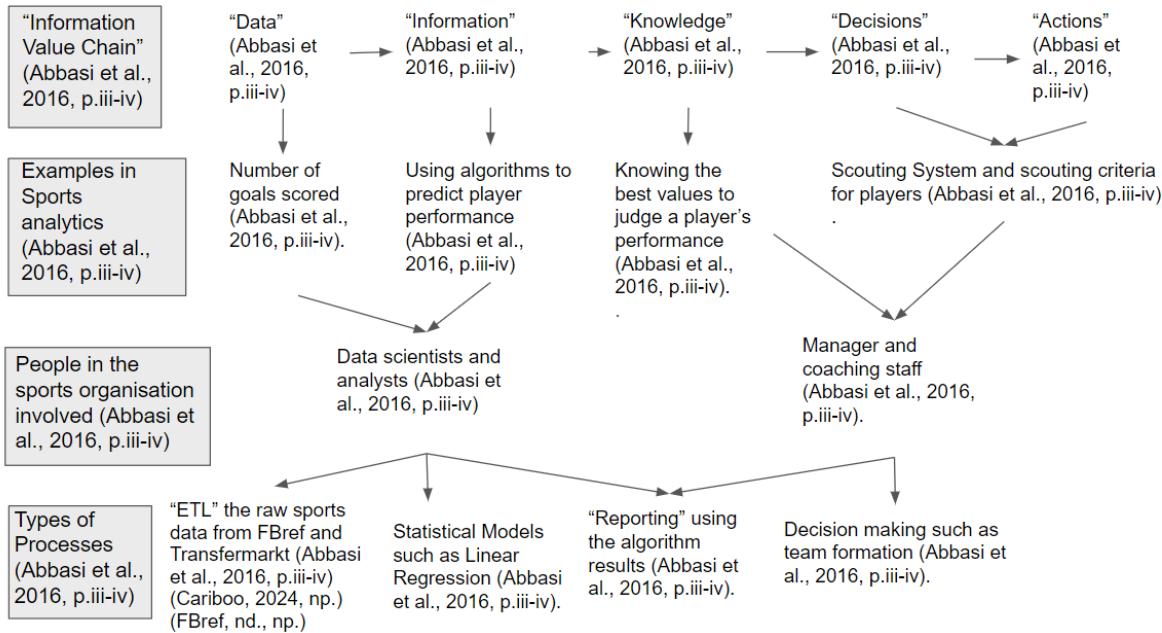
**Stage 4:** Data integration/building database (combining multiple data sources from different websites) (Saherwardi, 2019, np.) (Holland and Shaw, 2019). After preparing the data, the datasets can be “joined” together to make one big dataset with the relevant data. Integration will include taking “FBref” on-field data such as “goals” and combining them with “Transfermarkt” off-field data such as “estimated transfer value” (FBref, nd., np.) (Transfermarkt, nd., np.).

**Stage 5:** Data storage (structured data) (Saherwardi, 2019, np.) (Holland and Shaw, 2019). This project will store the data as structured data in CSV files (Microsoft, nd., np.).

**Stage 6:** Statistical and Machine Learning Applications (Saherwardi, 2019, np.) (Holland and Shaw, 2019). In the author's project, an algorithm that will be used is Random Forest Regression in order to predict a player's performance using the processed and ranked variables that Random Forest used to predict a player's performance (Beheshti, 2022, np.). Metrics such as mean squared error will be used to evaluate the accuracy of the algorithm (Beheshti, 2022, np.).

**Stage 7:** Understanding Results and Outcomes (Saherwardi, 2019, np.) (Holland and Shaw, 2019). Using the data, we can analyse, evaluate, and discuss the results and outcomes, along with the aid of visualisation tools such as Tableau (nd., np.). For example, the author could compare the rankings of variables that random forest used to predict player performances and how similar they are to their own Player Score, a composite statistic used as a general means to evaluate a player's overall performance relative to their position.

**Sports organisation “Information Value Chain” (Adapted from: Abbasi et al., 2016, p.iii-iv)**



From Data to Decisions: Developing a Data Science Approach To Football Using Moneyball Principles. IFIC

## **Chapter 4. Experiments, Evaluation, Results and Analysis**

### **4. Overview**

The two application areas chosen were: a scouting system designed to aid managers and their scouting staff by inputting desired criteria and outputting a list of players, allowing for more efficient scouting using data. Next, the RF algorithm used the dataset to train and predict players' performance and rank the most important metrics for predictions, giving management not only an improved selection of players and their performance but also undervalued metrics for scouting a player (Beheshti, 2022, np.).

All these applications remain in line with the Moneyball principles outlined in the methodology, as the management can use the applications as indicators of performance as well as input their desired transfer fees within their budget, allowing them to get undervalued or bargain players relative to their performance (Lewis, 2003).

#### **4.1 Scouting System**

The first managerial application that will be explored is the interactive scouting system.

The scouting system was made by importing the four cleaned and processed CSV datasets for each position that included the statistical analysis and predicted results. Using "widgets," an outfield player interface was created (figures 4.1.1-4.1.3 below), with an IntSlider being used as it will allow the user to customise the data and metrics to output the desired player (Juptyer Widgets, nd.). Another interface was created for goalkeepers using IntSlider (Juptyer Widgets, nd.). The code displayed can be updated to include other metrics, so despite not being included in the prototype scouting system displayed below, a metric such as "Ball Recoveries" can be included later.

Next, a function was created to allow the data to be filtered, such as when the slider is set to "5 goals," the output will only include those who scored "5 or more goals." This process was done for each selected metric, and this can be adjusted to a certain range, as displayed in the criteria for the IntSlider (Juptyer Widgets, nd.).

The figure below demonstrates an example of the interactive scouting system for the midfielder position. The sliders are all given a certain input, which generates an output of five midfielders that all match the desired inputs, with all the relevant data present. This allows a manager or scout to change the input based on their desired criteria for a player and give an output of players, allowing for further analysis.

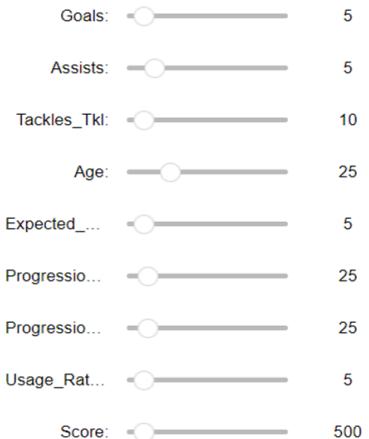
This process is then repeated for the three other positions, creating an interactive scouting system that a scout or manager can use to scout players with a desired criterion using the relevant data inputs.

## Figures 4.1.1-4.1.3

```
# Outfield Players' interface
Goals_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Performance_Gls:')
assists_slider = widgets.IntSlider(min=0, max=30, step=1, value=5, description='Performance_Ast:')
Tackles_slider = widgets.IntSlider(min=0, max=100, step=1, value=10, description='Tackles_Tkl:')
Age_slider = widgets.IntSlider(min=16, max=50, step=1, value=25, description='Age:')
npxG_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Expected_npxG:')
Progression_PrgC_slider = widgets.IntSlider(min=0, max=200, step=1, value=25, description='Progression_PrgC:')
Progression_PrgP_slider = widgets.IntSlider(min=0, max=200, step=1, value=25, description='Progression_PrgP:')
Usage_Rate_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Usage_Rate_Per:')
Score_slider = widgets.IntSlider(min=0, max=5000, step=1, value=500, description='Score:')
market_value_in_eur_slider = widgets.IntSlider(min=0, max=200000000, step=1000000, value=50000000, description='market_value_')
Tackles_TklW_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Tackles_TklW:')
Int_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Int:')
Blocks_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Blocks:')
Clr_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Clr:')
Challenges_Att_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Challenges_Att:')
Challenges_Tkl_per_slider = widgets.IntSlider(min=0, max=100, step=1, value=5, description='Challenges_Tkl_per:')
Challenges_Lost_slider = widgets.IntSlider(min=0, max=50, step=1, value=5, description='Challenges_Lost:')
Performance_CrdY_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Performance_CrdY:')
Performance_CrdR_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Performance_CrdR:')

# Goalkeepers' interface
Performance_Save_per_slider = widgets.IntSlider(min=0, max=100, step=1, value=0, description='Performance_Save_per:')
Penalties_saved_slider = widgets.IntSlider(min=0, max=20, step=1, value=0, description='Penalties_saved_Att:')
PSxG_minus_GA_slider = widgets.IntSlider(min=0, max=30, step=1, value=0, description='PSxG_minus_GA:')
Crosses_Stp_per_slider = widgets.IntSlider(min=0, max=30, step=1, value=0, description='Crosses_Stp_per:')
Number_of_Sweep_slider = widgets.IntSlider(min=0, max=10, step=1, value=0, description='Number_of_Sweep:')
Long_Pass_Acc_per_slider = widgets.IntSlider(min=0, max=50, step=1, value=0, description='Long_Pass_Acc_per:')
```

```
# Midfielders
# This function filters the dataset based on the slider values made by the user
# This will use the sliders for the outfield players and apply them to the midfielders
def filter_players_mid(Goals, Assists, Tackles_Tkl, Age, Expected_npxG, Progression_PrgC, Progression_PrgP, Usage_Rate_Per, Score):
    scout_interface = (mid_df['Goals'] >= Goals) &
                       (mid_df['Assists'] >= Assists) &
                       (mid_df['Tackles_Tkl'] >= Tackles_Tkl) &
                       (mid_df['Age'] <= Age) &
                       (mid_df['Expected_npxG'] >= Expected_npxG) &
                       (mid_df['Progression_PrgC'] >= Progression_PrgC) &
                       (mid_df['Progression_PrgP'] >= Progression_PrgP) &
                       (mid_df['Usage_Rate_Per'] >= Usage_Rate_Per) &
                       (mid_df['Score'] >= Score))
    )
    display(scout_interface)
```



Squad	Player	Fouls_Committed	Fouls_Won	Offsides	Crosses	Penalties_Won	Penalties_Conceded	Own_Goals	Ball_Recoveries	...	Height_in_C
3 Arsenal	Martin Ødegaard	30.0	36.0	3.0	77.0	1.0	0.0	0.0	158.0	...	178
7 RB Leipzig	Xavi Simons	32.0	70.0	1.0	76.0	2.0	0.0	0.0	167.0	...	179
12 Chelsea	Cole Palmer	23.0	26.0	6.0	81.0	1.0	0.0	0.0	105.0	...	189
28 Nott'ham Forest	Morgan Gibbs-White	37.0	57.0	6.0	182.0	0.0	0.0	0.0	179.0	...	171
30 Atalanta	Teun Koopmeiners	53.0	23.0	5.0	121.0	0.0	1.0	0.0	150.0	...	184

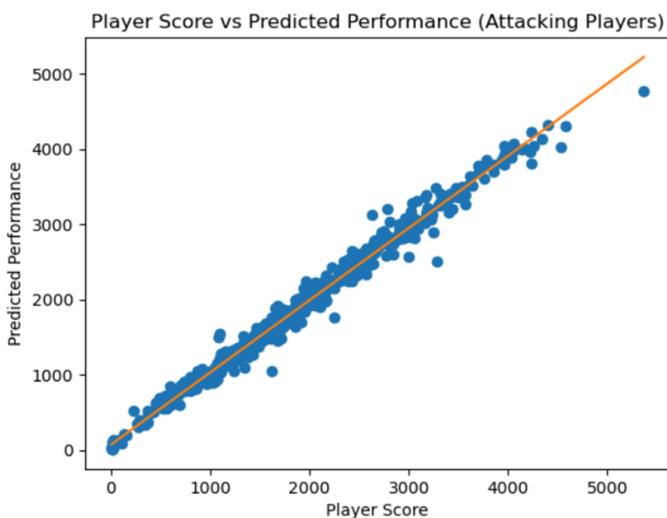
## 4.2 Player Score vs. Predicted Performance and Market Value vs Predicted Value Comparison

### Actual vs Predicted Player Score

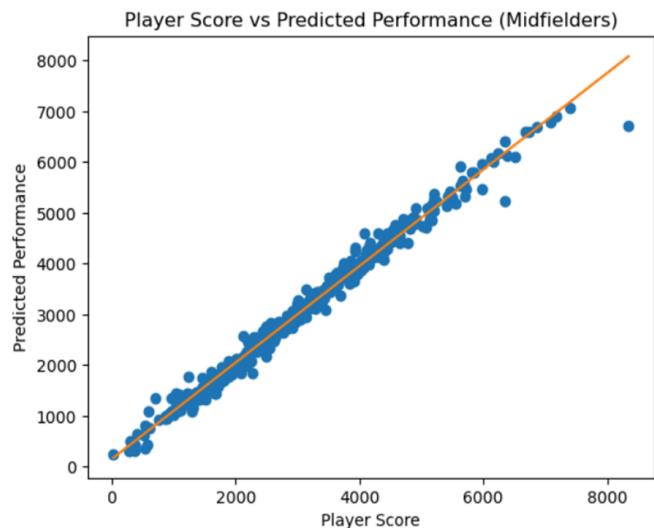
Here are initial results:

**Figures 4.2.1-4.2.4**

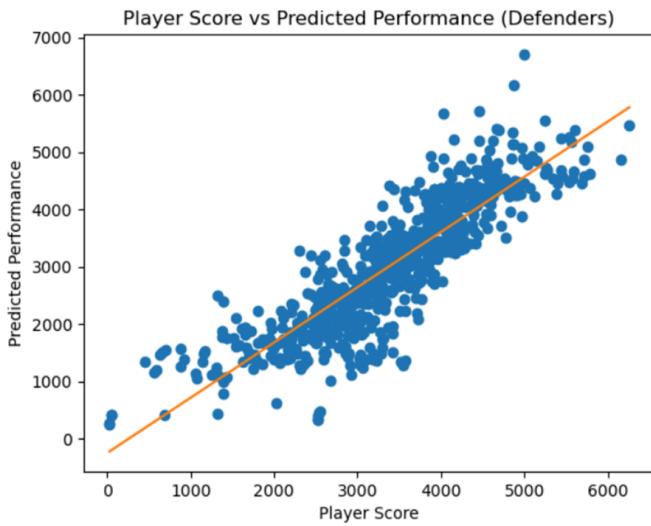
R-squared: 0.9868985587719865



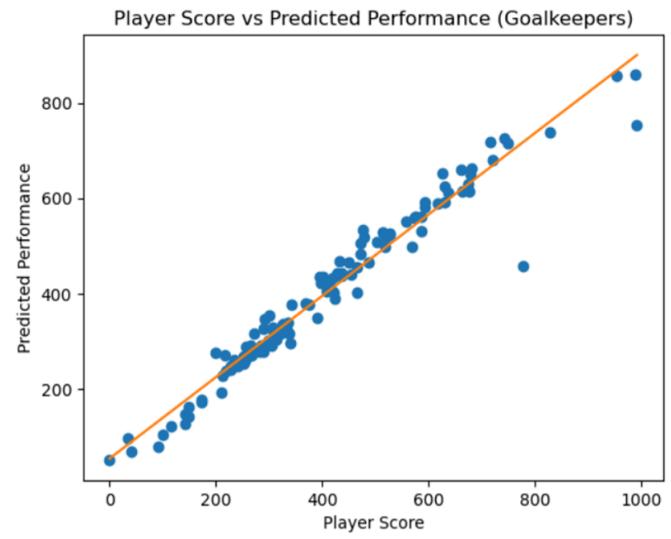
R-squared: 0.9865141285730494



R-squared: 0.7420941519921884



R-squared: 0.9531838742623864



**Table 4.2.1**

Position	MSE	MAE	RMSE	Normalised RMSE
Attacking Players	4)882199	4)786	4)939	4)0.04
Midfielders	2)98545	2)221	2)314	1)0.04
Defenders	3)485808	3)539	3)697	3)0.03
Goalkeepers	1)7776	1) 54	1)88	2)0.09

**Player Score vs Predicted Performance**

Now, the model displayed in the figures above was created by a program that compares the PS and PP.

This model uses the composite statistic “PS” as the variable that the algorithm would predict. The model displays both the linear relationship between the original PS and the predicted PS (“Predicted Performance”) by the algorithm and the accuracy of the prediction model through metrics such as MSE.

This model demonstrates the utility of composite statistics, as it can give a generally reliable and accurate indicator of a player’s performance, and when combined with a RF algorithm, it can provide sports management with a useful metric to measure a player’s overall performance and a ranking of the metrics that are most important to predicting a player’s performance, revealing some potentially undervalued metrics for scouting players (Beheshti, 2022, np.).

For the PS vs. PP results, we can observe that attacking players and midfielders had the highest correlation between their PS and their PP, followed by goalkeepers, and finally defenders. Meanwhile, for the normalised RMSE, the positions with the lowest RMSE (therefore, the positions with the least amount of error between the PS and PP) were defenders, followed by midfielders and attackers, and finally, with the highest normalised RMSE, goalkeepers.

For attacking players, an interesting observation is that there is a strong correlation between the PS and PP, but a low normalised RMSE. This could be because the PS and PP have nearly identical ratings for players, and the importance of the metrics to predict a player’s performance is similar. For example, in the PS equation for attacking players, there was a high weighting on goals, assists, progressive passes, carries, and take-on success, while the highly ranked features for PP also included playing time, passes received, and touches in the attacking third, which could explain why there is little error between the PS vs. PP results for attacking players.

Midfielders had a strong correlation and a low normalised RMSE, indicating there was little error between the PS and PP, and that a midfielder's score and their PP are almost completely aligned, demonstrating that in this scenario, RF was very effective (Beheshti, 2022, np.).

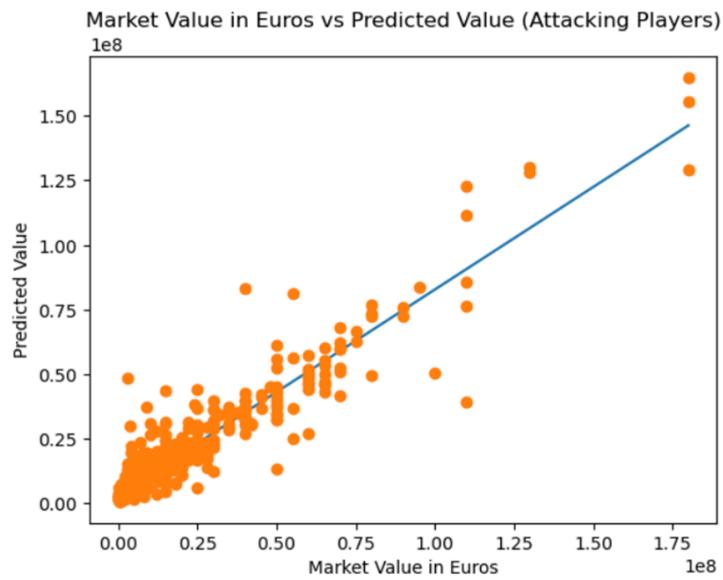
This is generally the same case as defenders had a strong, but in this case, weakest correlation as well as the lowest normalised RMSE. Both the algorithm and PS had an emphasis on tackles, interceptions, ball recoveries, etc., leading to a low amount of errors for the algorithm; however, the PP also had Playing Time metrics and Touch metrics ranked very highly, potentially explaining the difference in correlation between the PS and PP.

Finally, goalkeepers, with a strong correlation and a low normalised RMSE (but high relative to other positions (Bobbitt, 2021, np.). One of the higher-ranked variables, according to the algorithm, was “passes attempted,” which was not included in the PS function.

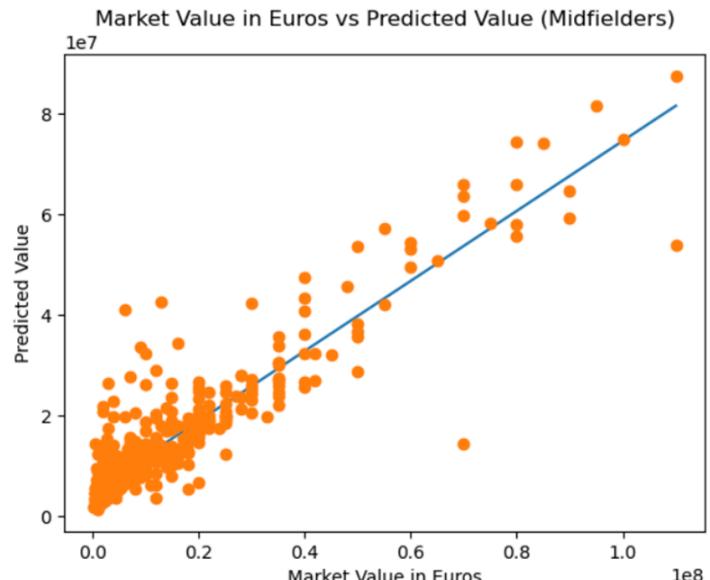
## 4.2 Market Value vs Predicted Value

**Figure 4.2.5-4.2.9**

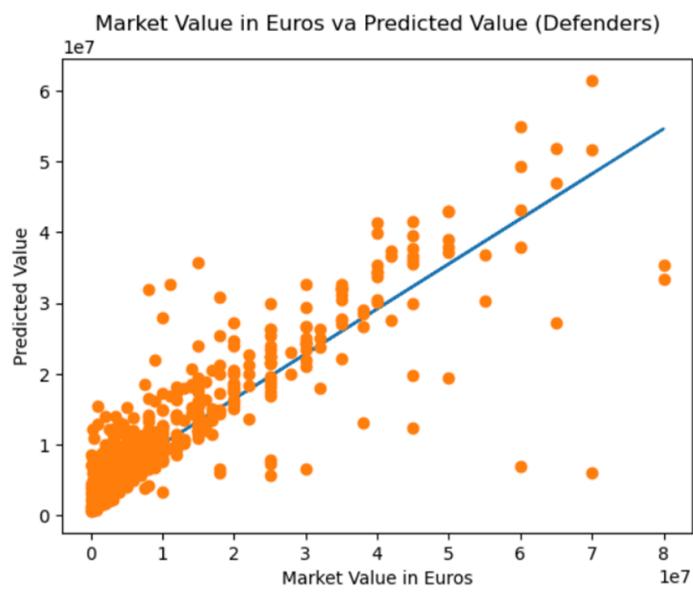
R-squared: 0.8879927563935



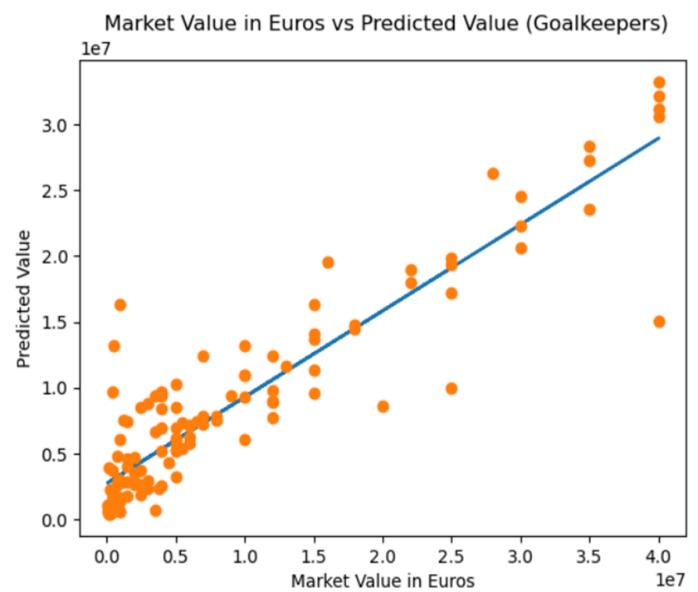
R-squared: 0.8358655336069594



R-squared: 0.7750000697992703



R-squared: 0.8340104468707997



**Table 4.2.2**

Position	MSE	MAE	RMSE	Normalised RMSE
Attacking Players	4)197790825073370	3)9309626	4)14063812	1)0.08
Midfielders	3)193532049707865	4)9960090	3)13911579	2)0.13
Defenders	2)170345637801076	2)7904020	2)13051652	3)0.16
Goalkeepers	1)67576926180288	1)5470278	1)8220518	4)0.21

Additionally, the project also created a program that predicts the PV of a footballer based on the “market value” variable.

This model uses the “market value” as the variable that the algorithm would predict, as the model displays both the linear relationship between the original Market Value and the predicted Market Value by the algorithm (“Predicted Value”) as well as the accuracy of the prediction model through metrics such as MSE.

This model aids the project's overall goal of finding undervalued players, as it can give a list of the player's PV, and when comparing the two values, one could compare the difference between the two to see if a player is undervalued or not. For example, if a player's market value is €40 million, but the PV is €60 million, then a player might be potentially undervalued by the market, as what the market and algorithm deem to make a player “valuable” might be different, as the algorithm uses over 80 variables to predict the value of a player, while human observers of the football transfer market who estimate a player's value might measure value differently, with their own personal bias involved, which will be discussed later in this section.

Firstly, looking at the initial results, one can observe that attacking players had the highest correlation between their market value and their PV followed by midfielders, then goalkeepers, and finally defenders. Meanwhile, for the normalised RMSE, the positions with the lowest RMSE (therefore, the positions with the least amount of error between the actual and PV) were attacking players, followed by midfielders, then defenders, and finally, with the highest normalised RMSE, goalkeepers (Beheshti, 2022, np.).

This could be because of what the market “values” in a player and what the algorithm ranks as the most important features. For example, the PV and Market Value of attacking players have a strong correlation, and there is little error between the actual and PV, which could be because both the market and algorithm both value similar metrics.

Now, back to what the “markets” deem to be important variables for a player’s value, the Franceschi et al. (2024) paper provides a “systematic review” of this topic as an overview of what determines a football player’s valuation. They use peer-reviewed, English-written papers on men’s football, selecting 29 papers with empirical models, to classify all the independent variables used to explain or predict a player’s valuation (Franceschi et al., 2024, pp.577-581). Franceschi et al. (2024, pp.580-581) also only selected papers where the dependent variable was either the player’s transfer fee or another publicly available and reliable valuation such as Transfermarkt, with 15 papers using Transfermarkt for their player valuations. Therefore, Franceschi et al.’s (2024, pp.580-581) paper results are a potentially good indicator of what Transfermarkt and the “football transfer market” believe are the variables that determine the value of a player.

The three metrics that the market used most to value a player were: number of goals and assists (or for a goalkeeper, the number of saves), since these metrics determine the winner of a football game; age (clubs will generally believe a 22-year-old who scores 25 goals has a higher ceiling and longevity than a 30-year-old who scores 25 goals); and play time, since naturally players who play more will show off their skills, have more eyes on them, and be noticed more often by the public (Franceschi et al., 2024, pp.590-597).

xG has also become valued more by clubs in recent years as a metric to determine a player’s “football IQ” or ability to get into attacking positions, so the author will also consider this important (Hewitt and Karakuş, 2023, pp.1-3).

Excluding the author’s own metrics, such as PS, the RF algorithm ranked the five most important features for an attacking player’s value to be: 1) Touches in the attacking penalty area 2) Touches in attacking third 3) Carries into the penalty box 4) Age; 5) Goals and Assists (Mafisher, 2023, np.).

The algorithm’s results are interesting as two of the three most valuable factors that determine an attacking player’s market value are present in the top five rankings but are not the most important, as well as both playing time and xG not ranking highly at all (FBref, nd., np.). However, it could be suggested that the top three variables are derivatives of the variables valued by the transfer market. For example, an attacking player who has a lot of touches in the attacking penalty area and attacking third of the pitch will generally get a lot of goals and assists, but even if they don’t, they can still be judged as a highly valuable asset to their team as they are still creating goal-scoring opportunities or increasing their team’s chances of scoring, which could also be highly valued by the transfer market.

Goalkeepers had the highest normalised RMSE, and defenders had the lowest correlation between their actual value and PV.

For goalkeepers, this could be because just making saves does not determine what actually makes a goalkeeper valuable. For example, a “save” in isolation does not elaborate on how good the save the keeper made was, which is where ‘Post-Shot-Expected-Goals Minus Goals

Allowed” comes in, as this determines the quality of the shot made at the keeper and whether they should have saved it or not, which is why according to the algorithm this was the most valuable metric for Goalkeeper and why potentially other metrics such as “crosses stopped” are undervalued by the market (FBref, nd. np.).

For defenders, this could also be the case as they do not score many goals or get as many assists as other outfield players, with many of their important metrics according to the algorithm being their tackles, clearances, blocks, etc., with defenders potentially being undervalued in the football market.

(See Appendix F for rankings of variables.).

## **Chapter 5. Discussion**

The scouting system could be useful for managers as it allows them to scout or identify a player with a desired criterion with relative ease. For example, say a Premier League manager wants a midfielder under the age of 25 who scores at least 5 goals, gets 5 assists. Using the scouting system and the interactive sliders, they input this desired criteria, giving them an output of players with the relevant data. Using the generated output, the manager can choose the player that fits these criteria. For example, using this criterion' Morgan Gibbs-White (market value is €40 million) is relatively cheap compared to the other options in the generated output (Martin Odegaard's market value is €95 million) and is already a Premier League player, demonstrating he is capable of high-level performances in the manager's club's league (Cariboo, 2024, np.).

This managerial application falls in line with the author's objective of creating a managerial application using Moneyball principles, since this low-cost, data-driven approach to scouting players can be altered and adjusted to the manager's desire (Lewis, 2003). For example, the scouting system could look for players with a high USP and a high number of ball recoveries and find the players that fit these criteria, creating an endless amount of adjustments and combinations for managers to find the desired players without having to invest in additional human scouts, potentially reducing costs and improving effectiveness when the scouting system and human scouts work together in cooperation (Lewis, 2023) (Raumdata, 2020, np.).

The managerial application of PS vs. PP includes demonstrating which variables are important indicators of a player's performance. The PS composite statistics are useful for general indicators of a player's performance relative to those in their position over a season, replicating the successful use of advanced statistics in basketball and the use of new metrics such as xG, which has been used by small clubs with "Moneyball Principles" such as Brighton and Brentford (Lewis, 2003) (Hewitt and Karakuş, 2023, pp.1-3).

Moreover, the RF algorithm also gave a ranking of the variables that affected a player's PP, indicating that scouts could potentially use these variable rankings, especially with positions with a low RMSE, to aid in scouting players by using undervalued metrics (Beheshti, 2022, np.). E.g., for midfielders, one of the three highest-ranked variables was progressive passes; therefore, when scouting midfielders, this should be one of the most valued metrics when judging their performance.

MSE, MAE, and RMSE are potentially high due to the "Curse of dimensionality," which is when there are numerous variables and a relatively small number of rows (or samples) (Adhikari, 2022, np.). This could be why goalkeepers have a lower MSE than the other positions due to having fewer variables and therefore fewer dimensions, leading to a lower MSE (Adhikari, 2022, np.).

Another reason that the Market Value vs. PV is so much higher than the PS vs. PP is because the transfer values are in the millions while the PS are in the thousands, so the "errors" will

inevitably be higher due to the greater magnitude of the data, with MSE being sensitive to outliers (Bobbitt, 2021, np.) (Beheshti, 2022, np.). Therefore, to get a better idea of the errors, the RMSE was normalised to get a better idea of which positions had the most accurate predictions (Bobbitt, 2021, np.).

However, despite the high number of metrics, they can still be used by comparing one another to figure out which position was the most accurately predicted by the model (Bobbitt, 2021, np.) (Beheshti, 2022, np.).

The RF algorithm proved to be very accurate, with the normalised RMSEs ranging from 0.03 to 0.21, with additional adjustment of the hyperparameter potentially leading to even better accuracy (Bobbitt, 2021, np.) (Beheshti, 2022, np.).

## **Chapter 6. Conclusions and Future Work**

Ultimately, the project objective was to contribute to the field of SA, specifically in football, as the literature review found that football was behind in the field of SA compared to basketball and baseball, with the project aiming to contribute to the literature through using ML, composite/advanced statistics, and programs with managerial applications (Szymanski, 2020, pp. 57-59).

The project succeeded in answering the research question by using the ML method, RF, to predict a player's performance and value, but also to rank the features by how they aided in predicting these results and to display which variables are potentially undervalued (Beheshti, 2022, np.). This leads into the next research question being answered, which is whether the project successfully used Moneyball principles such as composite and advanced statistics, undervalued variables, and linear regression for managerial applications such as the Attack/Defence Quadrant, the Moneyball Map, and the USP vs. PS graph (Raumdata, 2020, np.) (Lewis, 2003). Finally, the main SA program that was created for managerial purposes was the scouting system, which was an interactive system that allowed managers and their staff to put in input to generate the desired player output.

What makes this project stand out is its original scouting system, which was developed by the author with no observation or use of other scouting systems, as well as incorporating the author's own original composite statistics and advanced metric with the PS metric, which is also used in combination with the RF algorithm to improve scouting by predicting undervalued players and potentially discovering undervalued features in scouting players (Beheshti, 2022, np.).

This builds on the current literature, which has started to use SA methods such as ML to improve scouting and team selection. However, this project stands out as creating its own original advanced metric, which has been lacking in the football SA literature, and using it in combination with ML for regression, which has not been done in the literature so far, to create an original model that can be expanded upon in future research.

The research question of using Moneyball principles relates to the gap found in the literature review, as composite and advanced statistics were not as common in football as other sports such as basketball and baseball (Lewis, 2003). In addition, the author was unable to find the use of composite statistics and ML for regression in football analytics. Therefore, the project paired up composite statistics with ML and other forms of data analysis to improve scouting for a manager using a data-driven, or "Moneyball" approach (Lewis, 2003).

Moreover, the project also attempted to contribute to the literature with managerial applications such as the Attack/Defence quadrant and Moneyball map, as well as using the adapted USP from the basketball literature as a potential indicator of a player's performance level (Raumdata, 2020, np.) (Lewis, 2003). The author hopes that future research can be done using these applications and that they can be improved upon.

Some future work can also include expanding the datasets to include other football leagues, such as the Dutch and Portuguese leagues, since there is also a large amount of talent in these leagues with potential high-performing players that could go for lower transfer fees, upholding the idea of Moneyball (Lewis, 2003). These datasets that include these other leagues can be integrated into the project's dataset, allowing for an updated scouting system that includes players outside of Europe's top 5 leagues (FBref, nd., np.).

Finally, when using the RF algorithm, other algorithms such as Neural Networks could be used to predict player performance and value, and using metrics such as RMSE, their prediction accuracy could be compared to one another, with the most accurate ML method being used to predict the player's performance and value, potentially improving the results overall and leading to better outcomes for managers and their staff (Rodriguez-Galiano et al., 2015, p.804, p.811, p.816).

**Word count (excluding acknowledgements, title page, abstract, table of contents, bibliography, and appendices): 7,996.**

## Reference List

Abbasi, A., Sarker, S. and Chiang, R.H. (2016) Big data research in information systems: Toward an inclusive research agenda. *Journal of the association for information systems*, 17(2), pp.1-32. DOI: 10.17705/1jais.00423

APBRMetrics, [Online]. Available: <http://www.apbr.org/metrics/index.php>. In: Sarlis, V. and Tjortjis, C., (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, p.101562. DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

Adhikari, N. (2022) Curse of Dimensionality. *Medium* [online]. Available at: <https://medium.com/@nabinadadhikari190/curse-of-dimensionality-2d07f92bbf59> [Accessed on: 30th July 2024]

André, R. (2023) How to scrape and personalize data from FBref with Python: A guide to unlocking Football Insights. *Medium* [online]. Available at: <https://medium.com/@ricardoandreamo/how-to-scrape-and-personalize-data-from-fbref-with-python-a-guide-to-unlocking-football-insights-7e623607afca> [Accessed on: 5th June 2024]

Arkow, D., Jachuck, R., Kolar, L. (2023) MONEYB-BALL: An Analysis of Over and Undervalued NBA Statistics. *The Harvard College Sports Analysis Collective (HSAC)* [online]. Available at: <https://harvardsportsanalysis.org/2023/05/moneyb-ball-an-analysis-of-over-and-undervalued-nba-statistics/> [Accessed on: 15th July 2024]

Band, A (2020) Beginner’s guide to build Recommendation Engine in Python. *Medium* [online]. Available at: <https://medium.com/swlh/beginners-guide-to-build-recommendation-system-2bd4a96aa3e>

Bartosz, Ć., Giełczyk, A. and Choraś, M., (2021). Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, 23(1), p.90. DOI: <https://doi.org/10.3390/e23010090>

Basketball Reference (No date) Basketball Stats and History. *Basketball Reference* [online]. Available at:<https://www.basketball-reference.com/>

Basketball Reference (No date) Basketball Stats and History. *Basketball Reference* [online]. Available at: <https://www.basketball-reference.com/>  
In: Seshadri, R. (2004). “Improving Player Efficiency Rating in Basketball through Machine Learning.” *Research Archive of Rising Scholars*. Available at: <https://research-archive.org/index.php/rars/preprint/view/1000> DOI: <https://doi.org/10.58445/rars.1000>

Baumer, B., Zimbalist, A. (2014) Preface. The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball. *Philadelphia: University of Pennsylvania Press*, pp. ix-xiv. DOI: <https://doi.org/10.9783/9780812209129.ix>

Bharathiraja (2018) Random Forest Regression Using Python Sklearn From Scratch. *Medium* [online]. Available at: <https://medium.com/ampersand-academy/random-forest-regression-using-python-sklearn-from-scratch-9ad7cf2ec2bb> [Accessed on: 31st July 2024]

Bharadwaj, V. (2023) From Extinction to Excellence: Brighton's Data-Driven Triumph. *Medium*. Available at: <https://medium.com/@vishbs/from-extinction-to-excellence-brightons-data-driven-triumph-977abe738c1c> [Accessed on: 20th July 2024];

Beal, R., Norman, T.J. and Ramchurn, S.D., (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, 34, p.e28. DOI:10.1017/S0269888919000225

Beheshti, N. (2022) Random Forest Regression. *Medium* [online]. Available at: <https://towardsdatascience.com/random-forest-regression-5f605132d19d> [Accessed on: 31st July 2024]

Bobbitt, Z. (2021) What is Considered a Good RMSE Value? *Statology* [online]. Available at: <https://www.statology.org/what-is-a-good-rmse/> [Accessed on: 30th July 2024]

Bobbitt, Z. (2021) MSE vs. RMSE: Which Metric Should You Use? *Statology* [online]. Available at: <https://www.statology.org/mse-vs-rmse/> [Accessed on: 30th July 2024]

Campbell, J., and more (2022) What Premier League and EFL clubs really think of the Scottish market. *The Athletic*. Available at: <https://www.nytimes.com/athletic/3346810/2022/06/15/premier-league-epl-scottish-premiership-transfers/> [Accessed on: 10th August 2024]

Cariboo, D. (2024) 'Football Data from Transfermarkt'. *Kaggle* [online]. Available at: <https://www.kaggle.com/datasets/davidcariboo/player-scores/data> [Accessed on: 20th May 2024]

Chica-Olmo, M., Abarca, F., Rigol, J.P. (2002). Development of a decision support system based on remote sensing and GIS techniques for gold-rich area identification in SE Spain. *Int. J. Remote Sens.* 23, 4801–4814. In: Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R. (2015). Machine learning predictive models for mineral

prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, pp.804-818.

Chugh, A. (2020) MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium* [online]. Available at: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e> [Accessed on: 31st July 2024]

Collins, J. (2018) Michael Lewis's The Undoing Project: A Friendship That Changed The World. *Jason Collins Blog*. Available at: <https://www.jasoncollins.blog/posts/michael-lewiss-the-undoing-project-a-friendship-that-changed-the-world> [Accessed on: 17th July 2024]

Davenport, T.H., (2014). What businesses can learn from sports analytics. *MIT Sloan Management Review* [online]. Available at: <https://sloanreview.mit.edu/article/what-businesses-can-learn-from-sports-analytics/>

Depken, C.A. and Globan, T., (2021). Football transfer fee premiums and Europe's big five. *Southern economic journal*, 87(3), pp.889-908. DOI: 10.1002/soej.12471

Desmond, R. Explaining the Inverted Fullback—Player Role Analysis. (2022). Available online: <https://themastermindsite.com/2022/06/05/explaining-the-inverted-fullback-player-role-analysis/> [accessed on 13 May 2023]. In: Plakias, S., Moustakidis, S., Kokkotis, C., Papalex, M., Tsatalas, T., Giakas, G. and Tsaopoulos, D., (2023). Identifying soccer players' playing styles: a systematic review. *Journal of Functional Morphology and Kinesiology*, 8(3), p.104. DOI: <https://doi.org/10.3390/jfmk8030104>

Driblab (2022) xG Chain: the most important thing is to participate. *Driblab*. Available at: <https://www.driblab.com/analysis-player/xg-chain/> [Accessed on: 15th July 2024]

ESPN NBA stats, (2019) [Online]. Available: <https://www.espn.com/nba/stats>. [Accessed 20 August 2019] In: Sarlis, V. and Tjortjis, C., 2020. Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, p.101562.  
DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

Ferguson, R.B., (2014) B2B analytics in the airline market: Harnessing the power of consumer big data. *MIT Sloan Management Review*. Available at: <https://sloanreview.mit.edu/article/comscore-the-art-and-science-of-big-data-from-the-inside/> In: Holland, C.P., Thornton, S.C. and Naudé, P. (2020) B2B analytics in the airline market: Harnessing the power of consumer big data. *Industrial Marketing Management*, 86, pp.52-64.

FBref. <https://fbref.com/en/>. Cited in: Mead, J., O'Hare, A. and McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. Plos one, 18(4), p.e0282295.

FBref (No date) All About FBref.com. FBref [online]. Available at:  
<https://www.sports-reference.com/about.html>

FBref (No date) 2023-2024 Big 5 European Leagues: Advanced Goalkeeping. FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/keepersadv/players/2023-2024-Big-5-European-Leagues-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Goalkeeping. FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/keepers/players/2023-2024-Big-5-European-League-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Defensive Actions. FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/defense/players/2023-2024-Big-5-European-League-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Miscellaneous Stats. FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/misc/players/2023-2024-Big-5-European-Leagues-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Possession. FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/possession/players/2023-2024-Big-5-European-Leagues-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Possession (Squad). FBref [online]. Available at:  
<https://fbref.com/en/comps/Big5/2023-2024/possession/squads/2023-2024-Big-5-European-Leagues-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues: Standard Stats. FBref [online].

Available at:

<https://fbref.com/en/comps/Big5/2023-2024/stats/players/2023-2024-Big-5-European-Leagues-Stats> [Accessed on: 3rd June 2024]

FBref (No date) 2023-2024 Big 5 European Leagues. *FBref* [online]. Available at:

<https://fbref.com/en/comps/Big5/stats/squads/Big-5-European-Leagues-Stats>. In:

Raumdata (2020) Average Passing Distance and Usage Rate in the English Premier League.

*Raumdata*. Available at:

<https://raumdatafootball.com/2020/07/04/average-passing-distance-and-usage-rate-in-the-english-premier-league/> [Accessed on: 15th July 2024]

FIFA (2024) Latest Men's World Ranking. *FIFA* [online]. Available at:

<https://inside.fifa.com/fifa-world-ranking/men> [Accessed on: 27th July 2024]

Franceschi, M., Brocard, J.F., Follert, F. and Gouguet, J.J., (2024). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*, 38(3), pp.577-600. DOI: 10.1111/joes.12552

Gerrard, B (2017) *Chapter 14: Analytics, Technology and High Performance Sport*. In: Schulenkorf, N and Frawley, S, (eds.) *Critical Issues in Global Sport Management*. Routledge , London; New York . ISBN 9781138911222

Ghar, S., Patil, S. and Arunachalam, V., (2021), December. Data Driven football scouting assistance with simulated player performance extrapolation. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1160-1167). IEEE. DOI:[10.1109/ICMLA52953.2021.00189](https://doi.org/10.1109/ICMLA52953.2021.00189)

Hewitt, J.H. and Karakuş, O. (2023) A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, 4, 100034. pp.1-12. DOI: <https://doi.org/10.1016/j.fraope.2023.100034>

Hintz, E.S., (2022), April. Moneyball: The computational turn in professional sports management. In *Papers of the Business History Conference*. ISSN: 2573-6531

Holland, Christopher P., "Internet and Social Media Strategy in Sports Marketing" (2015). *ECIS* 2015 Completed Research Papers. Paper 79. ISBN 978-3-00-050284-2  
[http://aisel.aisnet.org/ecis2015\\_cr/79](http://aisel.aisnet.org/ecis2015_cr/79).

Holland, C.P., Thornton, S.C. and Naudé, P. (2020) B2B analytics in the airline market: Harnessing the power of consumer big data. *Industrial Marketing Management*, 86, pp.52-64. DOI: <https://doi.org/10.1016/j.indmarman.2019.11.002>

Hollinger, J. (2003). "Introducing PER." *Basketball Reference*. Available at: <https://tinyurl.com/bballRefpage>. In: Seshadri, R. (2004). "Improving Player Efficiency Rating in Basketball through Machine Learning." *Research Archive of Rising Scholars*. Available at: <https://research-archive.org/index.php/rars/preprint/view/1000> DOI: <https://doi.org/10.58445/rars.1000>

ipython (No date) Widget List. *ipython* [online]. Code Version: 6.0 Available at: "<https://ipython.readthedocs.io/en/stable/api/generated/IPython.display.html>" [Accessed on: 19th June 2024]

Jupyter (No date) About Us. *Jupyter* [online]. Available at: <https://jupyter.org/about>

Jupyter Widgets (No date) Widget List. *ipywidgets* [online]. Code Version: 7.6.2 Available at: "<https://ipywidgets.readthedocs.io/en/7.6.2/examples/Using%20Interact.html>" [Accessed on: 19th June 2024]

Jupyter Widgets (No date) Widget List. *ipywidgets* [online]. Code Version: 8.1.3 Available at: "<https://ipywidgets.readthedocs.io/en/latest/examples/Using%20Interact.html>" [Accessed on: 19th June 2024]

Karen, M. (2016) Five ways Arsenal's Arsene Wenger changed English football forever. *ESPN* [online]. Available at: [https://www.espn.co.uk/football/story/\\_/id/37456887/five-ways-arsenal-arsene-wenger-changed-english-football-forever](https://www.espn.co.uk/football/story/_/id/37456887/five-ways-arsenal-arsene-wenger-changed-english-football-forever) [Accessed on: 13th August 2024]

Kilvington, D., (2013). British Asians, covert racism and exclusion in English professional football. *Culture Unbound*, 5(4), pp.587-606. DOI:[10.3384/cu.2000.1525.135587](https://doi.org/10.3384/cu.2000.1525.135587)

Kinsella, N. (2024) Chelsea must 'trust' me to handle Cole Palmer's rise to superstardom, says Mauricio Pochettino. *The Standard* [online]. Available at: <https://www.standard.co.uk/sport/football/chelsea-fc-cole-palmer-mauricio-pochettino-b1152722.html> [Accessed on: 28th July 2024]

Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D.T. (2007) A starting point for analyzing basketball statistics. *Journal of quantitative analysis in sports*, 3(3). Pp.1-22. DOI:[10.2202/1559-0410.1070](https://doi.org/10.2202/1559-0410.1070)

Kumar, A. (2020) Random Forest for prediction. *Medium*. Available at:  
<https://towardsdatascience.com/random-forest-ca80e56224c1> [Accessed on: 31st July 2024]

Kuper, S (2024) Smart Money — Alex Duff on the changing fortunes of Brentford Football Club. *The Financial Times*. Available at:  
<https://www.ft.com/content/eb1ef9c8-cbe7-4080-8f1e-ec24bcd510ea> [Accessed on: 20th July 2024]

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York, NY: W. W. Norton & Company

Lewis, M., (2016). *The undoing project: A friendship that changed the world*. Penguin UK.

Lemire, J. (2023) Inside Brentford FC's key initiatives for continued EPL growth. *Sports Business Journal*. Available at:<https://www.sportsbusinessjournal.com/Articles/2023/08/02/brentford-fc#:~:text=Brentford%20in%20its%20third%20Premier,team%20and%20business%20strategy%20analytics>. [Accessed on: 20th July 2024]

Liu, A., Mahapatra, R.P. and Mayuri, A.V.R., (2023). Hybrid design for sports data visualization using AI and big data analytics. *Complex & Intelligent Systems*, 9(3), pp.2969-2980.  
<https://doi.org/10.1007/s40747-021-00557-w>

Machine learning Tutorials (2023) A Comprehensive Guide to the Pandas apply() Function in Python. *Machine learning Tutorials*. Available at:  
<https://machinelearningtutorials.org/a-comprehensive-guide-to-the-pandas-apply-function-in-python/> [Accessed on: 15th June 2024]

MaFisher (2023) Drafting The Perfect Prospect: Leveraging Data Science in MLB Draft Scenarios. *Medium* [online]. Available at:  
<https://mafisher.medium.com/drafting-the-perfect-prospect-leveraging-data-science-in-mlb-draft-scenarios-6444d9d7b89f> [Accessed on: 1st June 2024]

Matplotlib (No date) Title: Matplotlib 3.9.2 documentation. *Matplotlib* [online]. Code Version: 3.9.2. Available at: <https://matplotlib.org/stable/index.html> [Accessed on: 5th August 2024]

Mazari, N., (2024). Artificial Intelligence in Sports Training: A Reading on How AI Can Identify Sporting Talent in Football. University of Ziane Achour, ASJP, pp.08-20.

McDonnell, B., Sisneros, M., (2023) Brighton and Brentford: Two Smart Clubs Who Play the Game in Opposite Ways. *Opta Analyst*. Available

at:<https://theanalyst.com/eu/2023/03/brighton-and-brentford-two-smart-clubs-who-play-the-game-in-opposite-ways/> [Accessed on: 20th July 2024]

Microsoft (No date) Microsoft Excel. *Microsoft* [online]. Available at: <https://www.microsoft.com/en-us/microsoft-365/excel>

Mota, T., Silva, R. and Clemente, F., (2021). Holistic soccer profile by position: a theoretical framework. *Human movement*, 24(1), pp.4-20. doi: <https://doi.org/10.5114/hm.2023.110751>

Munzert, S, Rubba, C, Meissner, P and Nyhuis, D. (2014) Automated data collection with R: A practical guide to web scraping and text mining. *John Wiley & Sons*. DOI: <https://doi.org/10.1002/9781118834732> In: Thapelo, T.S., Namoshe, M., Matsebe, O., Motshegwa, T. and Bopape, M.J.M., 2021. SASSCAL WebSAPI: A Web Scraping Application Programming Interface to Support Access to SASSCAL's Weather Data. *Data Science Journal*, 20, pp.24-24. . DOI: <https://doi.org/10.5334/dsj-2021-024>

NBA basketball reference, (2019), [Online]. Available: <https://www.basketball-reference.com/>. [Accessed 20 August 2019] In: Sarlis, V. and Tjortjis, C., (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, p.101562. DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

NBA.com. [Online]. Available: <https://stats.nba.com> In: Sarlis, V. and Tjortjis, C., (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, p.101562. DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

numpy (No date) Title: NumPy user guide. *numpy* [online]. Code Version: 2.0. Available at: <https://numpy.org/doc/stable/user/index.html#user> [Accessed on: 5th August 2024]

OpenAI (2024) *ChatGPT* [online]. Available from: <https://chat.openai.com/> [Accessed June 3rd 2024].

Opta (No Date) Opta Analytics: AI-Derived Insights To Enhance Your Storytelling. *Opta* [online]. Available at: <https://football-observatory.com/IMG/sites/mr/mr25/en/> [Accessed on: 22nd July 2024]

Packer, J. (2018) Basketball and the Art of Non-Intuitive Decision Making. *Medium* [online]. Available at: <https://medium.com/@jpacker1/basketball-and-the-art-of-non-intuitive-decision-making-343da4d544b1> [Accessed on: 17th July 2024]

Pan, S.J. and Yang, Q., (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), pp.1345-1359. In: Beal, R., Norman, T.J. and Ramchurn, S.D., (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, 34, p.e28. DOI:10.1017/S0269888919000225

Pandas (No date) Title: User Guide - Pandas. *Pandas*. Available at:  
["https://pandas.pydata.org/docs/user\\_guide/index.html"](https://pandas.pydata.org/docs/user_guide/index.html) [Accessed on: 5th August 2024]

Piepenbreier, N (2023) How to Calculate R-Squared in Python (SkLearn and SciPy). *Datagy*. Available at: <https://datagy.io/python-r-squared/> [Accessed on: 16th July 2024]

Poli, R., Ravenel, L and Besson, R. (2017) 'CIES Football Observatory Monthly Report n°25 - May 2017. World expatriate footballers'. *CIES Football Observatory* [Online]. Available at: <https://football-observatory.com/IMG/sites/mr/mr25/en/> [Accessed on: 19th July 2024]

Plakias, S., Moustakidis, S., Kokkotis, C., Papalexis, M., Tsatalas, T., Giakas, G. and Tsaopoulos, D., (2023). Identifying soccer players' playing styles: a systematic review. *Journal of Functional Morphology and Kinesiology*, 8(3), p.104. DOI: <https://doi.org/10.3390/jfmk8030104>

Player Scout (No date) Football Academy FAQ's. Player Scout. Available at:  
[https://playerscout.co.uk/about-football-academies/#~:text=These%20academies%20sign%20several%20young,nationwide%20\(and%20sometimes%20internationally\).](https://playerscout.co.uk/about-football-academies/#~:text=These%20academies%20sign%20several%20young,nationwide%20(and%20sometimes%20internationally).)  
[Accessed on: 14th August 2024]

Price, S. (2023) 'Why More Brazilians Are On A Direct Journey To The Premier League'. *Forbes* [online]. Available at:  
<https://www.forbes.com/sites/steveprice/2023/01/24/why-more-brazilians-are-on-a-direct-journey-to-the-premier-league/> [Accessed on: 19th July 2024]

Puerzer, R.J., (2005). Baseball literary journals: The grass-roots literature of the game. *NINE: A Journal of Baseball History and Culture*, 14(1), pp.130-143. DOI: <https://doi.org/10.1353/nin.2005.0050>

Python (No date) About Python. *Python* [online]. Available at: <https://www.python.org/about/>

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(2proc), S514-S529.  
doi:<https://doi.org/10.14198/jhse.2017.12.Proc2.05>

Raumdata (2020) Average Passing Distance and Usage Rate in the English Premier League. *Raumdata* [online]. Available at: <https://raumdatafootball.com/2020/07/04/average-passing-distance-and-usage-rate-in-the-english-premier-league/> [Accessed on: 15th July 2024]

Research Archive of Rising Scholars (2024) About the Server. *Research Archive of Rising Scholars*. Available at: <https://research-archive.org/index.php/rars/about> Accessed on: [10th July 2024]

Rigol-Sanchez, J.P., Chica-Olmo, M., Abarca-Hernandez, F., (2003). Artificial neural networks as a tool for mineral potential mapping with GIS. *Int. J. Remote Sens.* 24, 1151–1156. In: Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, pp.804-818.

Robson, S. (2019) Why playing out from the back has brought mixed results for Premier League clubs. *ESPN* [online]. Available at: [https://www.espn.co.uk/football/story/\\_/id/37576665/why-playing-back-brought-mixed-results-premier-league-clubs](https://www.espn.co.uk/football/story/_/id/37576665/why-playing-back-brought-mixed-results-premier-league-clubs) [Accessed on: 25th July 2024]

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, pp.804-818. DOI: <https://doi.org/10.1016/j.oregeorev.2015.01.001>

Rytuba, J.J., Arribas Jr., A., Cunningham, C.G., McKee, E.H., Podwysocki, M.H., Smith, J.G., Kelly, W.C., Arribas, A., (1990). Mineralized and unmineralized calderas in Spain; part II, evolution of the Rodalquilar caldera complex and associated gold-alunite deposits. *Mineral Deposita* 25, S29–S35. In: Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, pp.804-818.

Saherwardi, I. A. (2023) Modern Data Architecture Concepts. *LinkedIn* [online]. Available at: <https://www.linkedin.com/pulse/modern-data-architecture-concepts-irfan-azim-saherwardi/>

Sapp, R.M., Spangenburg, E.E. and Hagberg, J.M., (2018). Trends in aggressive play and refereeing among the top five European soccer leagues. *Journal of sports sciences*, 36(12), pp.1346-1354. DOI: 10.1080/02640414.2017.1377911

Sarlis, V. and Tjortjis, C. (2020) Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, 101562. pp.1-19. DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

Schober, P., Boer, C. and Schwarte, L.A., (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5), pp.1763-1768. DOI: [10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864)

scikit-learn (No date) sklearn.model\_selection. *scikit-learn* [online]. Code Version: 1.5.1. Available at: "[https://scikit-learn.org/stable/api/sklearn.model\\_selection.html](https://scikit-learn.org/stable/api/sklearn.model_selection.html)" [Accessed on: 5th August 2024]

scikit-learn (No date) sklearn.ensemble. *scikit-learn* [online]. Code Version: 1.5.1. Available at: "<https://scikit-learn.org/stable/api/sklearn.ensemble.html>" [Accessed on: 5th August 2024]

scikit-learn (No date) sklearn.linear\_model. *scikit-learn* [online]. Code Version: 1.5.1. Available at: "[https://scikit-learn.org/stable/api/sklearn.linear\\_model.html](https://scikit-learn.org/stable/api/sklearn.linear_model.html)" [Accessed on: 5th August 2024]

scikit-learn (No date) sklearn.metrics. *scikit-learn* [online]. Code Version: 1.5.1. Available at: "<https://scikit-learn.org/stable/api/sklearn.metrics.html>" [Accessed on: 5th August 2024]

Seshadri, R. (2004). “Improving Player Efficiency Rating in Basketball through Machine Learning.” *Research Archive of Rising Scholars* [online]. Available at: <https://research-archive.org/index.php/rars/preprint/view/1000> DOI: <https://doi.org/10.58445/rars.1000>

Shaw and Holland (2019) Hierarchical Big data landscape Model. Research In Progress. In: Lecture notes and personal correspondence. Algorithms I. CP Holland, Loughborough University, (2024).

Shah, F.A., Kretzer, M. and Mädche, A. (2015) Designing an analytics platform for professional sports teams. *International Conference on Information Systems*. 1-18. At: Fort Worth, TX

Sofascore (<https://www.sofascore.com/>) cited in: Bartosz, Ć., Giełczyk, A. and Choraś, M., (2021). Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, 23(1), p.90. DOI: <https://doi.org/10.3390/e23010090>

Spencer, J. (2024) One-man team? Mauricio Pochettino admits his 'struggling' Chelsea players give the ball to Cole Palmer to 'solve' their problems. *GOAL* [online]. Available at: <https://goal.com/en-gb/lists/mauricio-pochettino-struggling-chelsea-players-give-ball-cole-palmer-solve-problems/bltb529dc8036ab5111#cscba4db3351083774> [Accessed on: 28th July 2024]

Spotrac (No date) NBA. *Spotrac* [online]. Available at: [https://www.spotrac.com/nba/cap/\\_year/2024/sort/cap\\_maximum\\_space](https://www.spotrac.com/nba/cap/_year/2024/sort/cap_maximum_space) In: Arkow, D., Jachuck, R., Kolar, L. (2023) MONEYB-BALL: An Analysis of Over and Undervalued NBA Statistics. *The Harvard College Sports Analysis Collective (HSAC)* [online]. Available at: <https://harvardsportsanalysis.org/2023/05/moneyb-ball-an-analysis-of-over-and-undervalued-nba-statistics/> [Accessed on: 15th July 2024]

Smith, R. (2021) The Wisdom of the Crowd. *The New York Times* [online]. Available at: <https://nytimes.com/2021/08/12/sports/soccer/soccer-football-transfermarkt.html> [Accessed on: 12th August 2024]

Srivastava, A., Chaudhary, A., Gupta, D. and Rana, A. (2021) Usage of Analytics in the World of Sports. In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-7). *IEEE*. <https://ieeexplore.ieee.org/abstract/document/9596466> DOI: 10.1109/ICRITO51393.2021.9596466

Statsbomb (No date) About StatsBomb. *Statsbomb* [online]. Available at: <https://statsbomb.com/who-we-are/> [Accessed on: 14th July 2024] In: Hewitt, J.H. and Karakuş, O., (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, 4, p.100034. DOI: <https://doi.org/10.1016/j.fraope.2023.100034>

Statsbomb (No date) About StatsBomb. *Statsbomb* [online]. Available at: <https://statsbomb.com/who-we-are/> [Accessed on: 14th July 2024] In: Mead, J., O'Hare, A. and McMenemy, P., 2023. Expected goals in football: Improving model performance and demonstrating value. *Plos one*, 18(4), p.e0282295.

Stats Perform (No date) Live Tracking Data to Power Your In-Game & Post Match Analysis. *Stats Perform*. Available at: <https://www.statsperform.com/team-performance/football-performance/sportvu/> In: Sarlis, V. and Tjortjis, C., 2020. Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, p.101562. DOI:[10.1016/j.is.2020.101562](https://doi.org/10.1016/j.is.2020.101562)

St. John, A. (2013) Powered By Bill James And Friends, The Red Sox Win (Another) Moneyball World Series. *Forbes* [online]. Available at:

<https://www.forbes.com/sites/allenstjohn/2013/10/31/powerd-by-bill-james-and-friends-the-red-sox-win-another-moneyball-world-series/> [Accessed on: 23rd July 2024]

Szymanski, S. (2020). Sport analytics: Science or alchemy?. *Kinesiology Review*, 9(1), pp.57-63. DOI: <https://doi.org/10.1123/kr.2019-0066>

Tableau (No Date) What is Tableau? *Tableau* [online]. Available at:  
<https://www.tableau.com/en-gb/why-tableau/what-is-tableau> [Accessed on: 31st July 2024]

Taylor, M. (2006). Global players? Football, migration and globalization, c. 1930-2000. *Historical Social Research*, 31(1), 7-30. <https://doi.org/10.12759/hsr.31.2006.1.7-30>

Terner, Z. and Franks, A., (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, 8(1), pp.1-23. Available at:  
<https://arxiv.org/pdf/2007.10550.pdf>

Thaler, R.H. and Sunstein, C.R., (2003). Market efficiency and rationality: The peculiar case of baseball. *Mich. L. Rev.*, 102, p.1390-1403. Available at:  
<https://repository.law.umich.edu/mlr/vol102/iss6/18>

The Football Analyst (2024) Atalanta - Gian Piero Gasperini - Tactical Analysis. *The Football Analyst* [online]. Available at:  
<https://the-footballanalyst.com/atalanta-gian-piero-gasperini-tactical-analysis/>  
[Accessed on: 28th July 2024]

The Football Analyst (2024) Inter Milan - Simone Inzaghi - Tactical Analysis. *The Football Analyst* [online]. Available at:  
<https://the-footballanalyst.com/inter-milan-simone-inzaghi-tactical-analysis/>  
[Accessed on: 28th July 2024]

Thakkar, P. and Shah, M. (2021) An assessment of football through the lens of data science. *Annals of Data Science*, pp.1-14. DOI: 10.1007/s40745-021-00323-2

Tharme, L. (2024) Kylian Mbappe to Real Madrid makes more tactical sense than ever - this is why. *The Athletic* [online]. Available at:  
<https://www.nytimes.com/athletic/5287291/2024/06/04/kylian-mbappe-real-madrid-psg-ancelotti-tactics/> [Accessed on: 28th July 2024]

Transfermarkt. <https://www.transfermarkt.co.uk/>. Cited in:

Bartosz, Ć., Giełczyk, A. and Choraś, M., (2021). Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, 23(1), p.90. DOI: <https://doi.org/10.3390/e23010090>

Transfermarkt. <https://www.transfermarkt.co.uk/>. Cited in:  
Mead, J., O'Hare, A. and McMenemy, P., (2023). Expected goals in football: Improving model performance and demonstrating value. *Plos one*, 18(4), p.e0282295.

Transfermarkt.com (No date) Spotlights. *Transfermarkt* [online]. Available at: <https://www.transfermarkt.com/>

Troilo, M., Bouchet, A., Urban, T.L. and Sutton, W.A. (2016) Perception, reality, and the adoption of business analytics: Evidence from North American professional sport organizations. *Omega*, 59, pp.72-83. DOI: 10.1016/j.omega.2015.05.011

Tversky, Amos, and Daniel Kahneman. "Judgement under Uncertainty: Heuristics." *Science* 185 (1974): 1124-1131. In: Berri, D.J. and Eschker, E., (2005). Performance when it counts? The myth of the prime time performer in professional basketball. *Journal of Economic Issues*, 39(3), pp.798-807. DOI: 10.1080/00213624.2005.11506847

Tweedale, A., Clinard, B., Manuel, J. (2023) Pedro Porro is the Perfect Emblem of Ange Postecoglou's Success at Spurs. *Opta Analyst* [online]. Available at: <https://theanalyst.com/eu/2023/12/pedro-porro-ange-postecoglou-spurs-inverted-full-back/> [Accessed on: 28th July 2024]

UEFA (No date) UEFA rankings. *UEFA* [online]. Available at: <https://www.uefa.com/nationalassociations/uefarankings/country/?year=2025> [Accessed on: 27th July]

Ungruhe, C. and Schmidt, M.B., (2020). Why are East African players absent in European football? Localizing African football migration along structural constraints, colonial legacies and voluntary immobility. *Journal of Sport and Social Issues*, 44(5), pp.397-420. DOI: <https://doi.org/10.1177/0193723520919820>

Van Deusen, A (2023) Python Popularity: The Rise of A Global Programming Language. *Flatiron School* [online]. Available at: <https://flatironschool.com/blog/python-popularity-the-rise-of-a-global-programming-language/>

WhoScored (<https://www.whoscored.com/>) in: Bartosz, Ć., Giełczyk, A. and Choraś, M., (2021). Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, 23(1), p.90. DOI: <https://doi.org/10.3390/e23010090>

Wixom B.H., Ross, J.W., Beath, C.M., Miller, C.A. (2013) Capturing value from big data at comscore through platform, people, and perception. *MIT CISR*. In: Holland, C.P., Thornton, S.C. and Naudé, P., (2020). B2B analytics in the airline market: Harnessing the power of consumer big data. *Industrial Marketing Management*, 86, pp.52-64. DOI: <https://doi.org/10.1016/j.indmarman.2019.11.002>

Wyscout (No date) ‘Glossary’. *Wyscout* [online]. Available at: <https://dataglossary.wyscout.com/duel/#:~:text=A%20challenge%20between%20two%20players,ball%20or%20change%20its%20direction>. [Accessed on: 22nd July 2024]

Wyscout (2020) Soccer match event dataset. *Wyscout*. available at: [https://figshare.com/collections/Soccer\\_match\\_event\\_dataset/4415000/5](https://figshare.com/collections/Soccer_match_event_dataset/4415000/5) -. Cited in: Mead, J., O’Hare, A. and McMenemy, P., (2023). Expected goals in football: Improving model performance and demonstrating value. *Plos one*, 18(4), p.e0282295.

Xerxes (no date) MARKET VALUE DEFINITION. *Transfermarkt* [online]. Available at: [https://www.transfermarkt.co.uk/market-value-definition/thread/forum/357/thread\\_id/3433#:~:text=In%20general%2C%20the%20Transfermarkt%20market,relevant%20in%20determining%20market%20values](https://www.transfermarkt.co.uk/market-value-definition/thread/forum/357/thread_id/3433#:~:text=In%20general%2C%20the%20Transfermarkt%20market,relevant%20in%20determining%20market%20values).

## Appendices

### Appendix A

#### Literature review table

Paper	Topic(s)	Data Source	Methods/Data
Lewis, M. (2003). Moneyball: The art of winning an unfair game. New York, NY: W. W. Norton & Company	In this book, Lewis (2003) tracks the use of data analytics by the Oakland Athletics, one of the poorest baseball franchises, and how they used new digital technology to give themselves a competitive advantage by scouting players that were undervalued by traditional scouting and became one of the best baseball franchises over this period (Lewis, 2003).	N/A	Use of advanced baseball metrics (Sabermetrics) and basic machine learning tools such as linear regression (Lewis, 2003).
Srivastava, A., Chaudhary, A., Gupta, D. and Rana, A. (2021) Usage of Analytics in the World of Sports. In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-7). IEEE. <a href="https://ieeexplore.ieee.org/abstract/document/9596466">https://ieeexplore.ieee.org/abstract/document/9596466</a>	A useful introductory text that explains the fundamentals of sports analytics, including on-field vs. off-field analytics, supervised vs. unsupervised machine learning, and the use of techniques such as data visualisation and how they can be used for purposes such as game prediction, strategic game decisions, etc. (Srivastava, 2021, pp.1-4).	N/A	N/A
Hintz, E.S., 2022, April. Moneyball: The computational turn in professional sports management. In Papers of the Business History Conference.	Provides a brief history of sports analytics, including the rise of the “analytics movement” and how it was eventually embraced by management (Hintz, 2022, pp.1-7). Moreover, it also provides a process on how to create a suitable sports analytical data model (Hintz, 2022, pp.8-12).	N/A	N/A
Beal, R., Norman, T.J. and Ramchurn, S.D., 2019. Artificial intelligence for team sports: a survey. The Knowledge Engineering Review, 34, p.e28.DOI: <a href="https://doi.org/10.1017/S0269888919000225">10.1017/S0269888919000225</a>	The paper covers the current use of AI and ML in sports for “match outcome prediction, tactics, player transfers, fantasy sports, and predicting injuries” (Beal et al., 2019, pp.12-20).  The use of these data science techniques was done to help sports teams improve their strategic and tactical decision-making (Beal et al., 2019, pp.12-20).  Additionally, the paper discusses the strengths and weaknesses of the ML models and techniques that have been used (Beal et al., 2019, pp.12-20).	N/A	Discusses the use of ML methods: “neural networks, decision trees, clustering methods, SVM, random forest, and gradient boosting” (Beal et al., 2019, pp.10-20).  Other methods used: “Deep imitation learning, team

			formation optimisation, deep learning, Recurrent neural networks, Bayesian approaches, least squares and decision trees, and mixed-integer programming” (Beal et al., 2019, pp.10-20).
Mazari, N., 2024. Artificial Intelligence in Sports Training: A Reading on How AI Can Identify Sporting Talent in Football. University of Ziane Achour, ASJP, pp.08-20.	The paper provides a useful explanation of terminology relating to sports analytics and gives modern updates on the use of AI in sports analytics (Mazari, 2024, pp.15-17).	N/A	N/A
Seshadri, R. (2004). “Improving Player Efficiency Rating in Basketball through Machine Learning.” Research Archive of Rising Scholars. Available at: <a href="https://research-archive.org/index.php/rars/preprint/view/1000">https://research-archive.org/index.php/rars/preprint/view/1000</a> DOI: <a href="https://doi.org/10.58445/rars.1000">https://doi.org/10.58445/rars.1000</a>	Use of composite/advanced statistics and AI/ML to help scout and evaluate players (Seshadri, 2004, p.1-3).	Basketball Reference (nd., cited in: Seshadri, 2004, p.1-3)	Web scraping the website “Basketball Reference” (Seshadri, 2004, p.1-3).  Algorithms used included: Neural Networks (PyTorch), LASSO Regression, and Random Forest Regression (Seshadri, 2004, p.1-3).
Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D.T. (2007) A starting point for analyzing basketball statistics. Journal of quantitative analysis in sports, 3(3). Pp.1-22. DOI: <a href="https://doi.org/10.2202/1559-0410.1070">10.2202/1559-0410.1070</a>	These 3 articles cover the concept of advanced statistics in basketball and how they can be used to measure a player’s performance, improve a coach’s ability to select players, improve scouting, etc., with Sarlis and Tjortjis (2020, pp.1-2) examining how ML algorithms can be used to improve these advanced metrics.	Basketball-reference.com (2019, cited in: Sarlis and Tjortjis, 2020, pp.1-5, pp.17-18), SportVU (Stats Perform, nd., cited in: Sarlis and Tjortjis, 2020, pp.1-5, pp.17-18), NBA.com (nd., cited in: Sarlis and Tjortjis, 2020, pp.5-18), ESPN.com (2019 cited in: Sarlis and Tjortjis, 2020, pp.1-5, pp.17-18), APBRmetrics Forum (2019 cited in: Sarlis and Tjortjis, 2020, pp.1-5, pp.17-1)	Advanced metrics, statistical analysis and ML algorithms such as “Association Rule based models” (Sarlis and Tjortjis, 2020, pp.1-2).

Its Application, 8(1), pp.1-23. Available at: <a href="https://arxiv.org/pdf/2007.10550">https://arxiv.org/pdf/2007.10550</a>			
Franks, A., Miller, A., Bornn, L. and Goldsberry, K., 2015, February. Counterpoints: Advanced defensive metrics for nba basketball. In 9th annual MIT sloan sports analytics conference, Boston, MA (Vol. 10).  <a href="http://www.lukebornn.com/papers/franks_ssac_2015.pdf">http://www.lukebornn.com/papers/franks_ssac_2015.pdf</a>	<p>This article demonstrates basketball's advanced development in sports analytics relative to football (Franks et al., 2015, pp.1-6). As they continue to improve the metrics by recognising that offensive metrics are more developed, defensive metrics also need to be improved (Franks et al., 2015, pp.1-6).</p> <p>The article is also helpful, as it gives some insight into how to develop advanced or new metrics (Franks et al., 2015, pp.1-6).</p> <p>The article creates a model that estimates defensive match-ups for an entire basketball game and presents multiple metrics for evaluating a player's defensive performance (Franks et al., 2015, pp.1-6).</p>	<p>Tracking data from the 2013-2014 NBA season (does not reveal source but is probably Basketball-Reference.com, cited in: Franks et al., 2015, pp.1-6).</p>	<p>Markov model, weighted least squares, and expectation-maximisation (Franks et al., 2015, pp.1-6).</p>
Lewis, M., 2016. The undoing project: A friendship that changed the world. Penguin UK.  Thaler, R.H. and Sunstein, C.R., 2003. Market efficiency and rationality: The peculiar case of baseball. <i>Mich. L. Rev.</i> , 102, p.1390-1403.  Tversky, Amos, and Daniel Kahneman. "Judgement under Uncertainty: Heuristics." <i>Science</i> 185 (1974): 1124-1131. In: Berri, D.J. and Eschker, E., 2005. Performance when it counts? The myth of the prime time performer in professional basketball. <i>Journal of Economic Issues</i> , 39(3), pp.798-807.  Collins, J. (2018) Michael Lewis's The Undoing Project: A Friendship That Changed The World. Jason Collins Blog. Available at: <a href="https://www.jasoncollins.blog/post/the-undoing-project-a-friendship-that-changed-the-world">https://www.jasoncollins.blog/post/the-undoing-project-a-friendship-that-changed-the-world</a>	<p>The Undoing Project by Lewis (2016) and supplemented by articles related to this book and Moneyball (Lewis, 2003) explains how human error and mistakes, work that was pioneered by Tversky and Kahneman (1974, cited in: Berri and Eschker, 2005, pp. 799–800), led to market inefficiencies and underperformance in many sectors of the economy, with Lewis (2016) providing the example of the sports transfer markets, specifically basketball, with human errors leading to Daryl Morey ignoring his own data-based approach, leading to players that turned out to be successful being overlooked (Collins, 2018, np.) (Packer, 2018, np.) (Thaler and Sunstein, 2004, pp.1390-1398).</p> <p>The literature's main idea can be summarised as human error leading to irrational decisions and that a data-driven approach can lead to fewer inefficiencies and better informed decisions (Collins, 2018, np.) (Lewis, 2016, pp.13-49) (Packer, 2018, np.) (Thaler and Sunstein, 2004, pp.1390-1398) (Tversky and Kahneman, cited in: Berri and Eschker, 2005, pp.799-800).</p>	N/A	N/A

<p><a href="https://medium.com/@jpacker1/basketball-and-the-art-of-non-intuitive-decision-making-343da4d544b1">ts/michael-lewiss-the-undoing-project-a-friendship-that-changed-the-world</a> [Accessed on: 17th July 2024]</p> <p>Packer, J. (2018) Basketball and the Art of Non-Intuitive Decision Making. Medium. Available at: <a href="https://medium.com/@jpacker1/basketball-and-the-art-of-non-intuitive-decision-making-343da4d544b1">https://medium.com/@jpacker1/basketball-and-the-art-of-non-intuitive-decision-making-343da4d544b1</a> [Accessed on: 17th July 2024]</p>			
<p>Hewitt, J.H. and Karakuş, O., 2023. A machine learning approach for player and position adjusted expected goals in football (soccer). Franklin Open, 4, p.100034. DOI: <a href="https://doi.org/10.1016/j.fraope.2023.100034">https://doi.org/10.1016/j.fraope.2023.100034</a></p>	<p>The article continues to develop the xG metric by including more features than are used in other xG models (Hewitt and Karakuş, 2023, pp.1-3). The features used are “Goalkeeper positioning, Player Pressure radiiuses, and opposition between the shot and the goal.” (Hewitt and Karakuş, 2023, pp.1-3).</p> <p>The paper does a fantastic job explaining xG and applying ML methods to improve it (Hewitt and Karakuş, 2023, pp.1-3).</p>	<p>Statsbomb (nd., Hewitt and Karakuş, 2023)</p>	<p>Decision Trees, Gradient Boosting (Hewitt and Karakuş, 2023, pp.3-8)</p>
<p>Raumdata (2020) Average Passing Distance and Usage Rate in the English Premier League. Raumdata. Available at: <a href="https://raumdatafootball.com/2020/07/04/average-passing-distance-and-usage-rate-in-the-english-premier-league/">https://raumdatafootball.com/2020/07/04/average-passing-distance-and-usage-rate-in-the-english-premier-league/</a> [Accessed on: 15th July 2024]</p>	<p>Introduces the basketball metric of “usage rate %” to football, which is calculated by taking the number of touches of the ball a player has in the game divided by the number of touches the rest of the team has as a way to determine “how much of the ball” a player has (Raumdata, 2020, np.).</p>	<p>FBref (nd., cited in: Raumdata, 2020).</p>	<p>“Usage Rate Percentage” metric (Raumdata, 2020, np.).</p>
<p>Kuper, S (2024) Smart Money — Alex Duff on the changing fortunes of Brentford Football Club. The Financial Times. Available at: <a href="https://www.ft.com/content/eb1ef9c8-cbe7-4080-8f1e-ec24bcd510ea">https://www.ft.com/content/eb1ef9c8-cbe7-4080-8f1e-ec24bcd510ea</a> [Accessed on: 20th July 2024]; McDonnell, B., Sisneros, M., (2023) Brighton and Brentford: Two Smart Clubs Who Play the Game in Opposite Ways. Opta Analyst. Available at:<a href="https://theanalyst.com/eu/2023/03/brighton-and-brentford-two-smart-clubs-who-play-the-game-in">https://theanalyst.com/eu/2023/03/brighton-and-brentford-two-smart-clubs-who-play-the-game-in</a></p>	<p>These articles have been grouped together as they cover the rise of the clubs of Brentford and Brighton.</p> <p>Kuper (2024) and Lemire (2023) focus on the rise of Brentford FC through its use of data analytics and statistical analysis to buy undervalued players, sell them for a profit, and improve their tactics and team selection. Lemire (2023, np.) also discusses how Brentford are now using their data analytics for off-field purposes such as business decisions and improving the match-going fan experience.</p>	<p>Opta Analyst (McDonnell and Sisneros, 2023).</p>	<p>N/A</p>

<p><a href="#">_opposite-ways/</a> [Accessed on: 20th July 2024];</p> <p>Bharadwaj, V. (2023) From Extinction to Excellence: Brighton's Data-Driven Triumph. Medium. Available at:<a href="https://medium.com/@vishbs/rom-extinction-to-excellence-brughtons-data-driven-triumph-977ab-e738c1c">https://medium.com/@vishbs/rom-extinction-to-excellence-brughtons-data-driven-triumph-977ab-e738c1c</a> [Accessed on: 20th July 2024];</p> <p>Lemire, J. (2023) Inside Brentford FC's key initiatives for continued EPL growth. Sports Business Journal. Available at:<a href="https://www.sportsbusinessjournal.com/Articles/2023/08/02/brentford-fc#:~:text=Brentford%2C%20in%20its%20third%20Premier,team%20and%20business%20strategy%20analytics">https://www.sportsbusinessjournal.com/Articles/2023/08/02/brentford-fc#:~:text=Brentford%2C%20in%20its%20third%20Premier,team%20and%20business%20strategy%20analytics</a>. [Accessed on: 20th July 2024]</p>	<p>Lemire (2023) focusses on Brighton's rise through the use of data analytics, being one of the first football clubs to use expected goals in their scouting methods. Like Brentford, they scout undervalued players who perform well for the club and are usually sold for a massive profit (Lemire, 2023).</p> <p>McDonnell and Sisneros (2023) focus more on how the clubs use of data analytics influences their style of player and how, using these analytics, they can scout and select players that fit their system. They then go into detail about how each club plays a specific brand of football, with Brighton preferring attacking football and Brentford preferring defensive football (McDonnell and Sisneros, 2023).</p>		
<p>Kumar, A. (2020) Random Forest for prediction. Medium. Available at: <a href="https://towardsdatascience.com/random-forest-ca80e56224c1">https://towardsdatascience.com/random-forest-ca80e56224c1</a> [Accessed on: 31st July 2024]</p> <p>Beheshti, N. (2022) Random Forest Regression. Medium. Available at: <a href="https://towardsdatascience.com/random-forest-regression-5f605132d19d">https://towardsdatascience.com/random-forest-regression-5f605132d19d</a> [Accessed on: 31st July 2024]</p> <p>Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, pp.804-818.</p>	<p>These articles provide a good summary of the Random Forest algorithm, how to measure its accuracy through the use of metrics such as MSE, and the pros and cons of the algorithm overall (Kumar, 2020, np.) (Beheshti, 2022, np.) (Rodriguez-Galiano et al., 2015, p.804, p.816).</p> <p>To examine the accuracy of random forest, the author looked at a paper examining machine learning models being used on “mineral prospectivity,” with the random forest algorithm being found to be most accurate (Rodriguez-Galiano et al., 2015, p.804, p.816).</p>	<p>Rigol-Sanchez et al. (2003) and Chica-Olmo et al. (2002) integrated all the datasets, facilitated by ADARO, S.A. and collected during the DARSTIMEX Project (University of Granada) with additional data from Rytuba et al. (1990) (cited in: Rodriguez-Galiano et al., 2015, p.804, p.816))</p>	<p>Random Forest, neural networks, random forest, regression trees, and support vector machines (Rodriguez-Galiano et al., 2015, p.804, p.816).</p> <p>The database is made up of “46 gold occurrence locations” (Rodriguez-Galiano et al., 2015, p.804, p.816).</p> <p>The data also includes “exploited deposits” and “mineralised structures” (Rodriguez-Galiano et al., 2015, p.804, p.816).</p> <p>Physical and chemical data: “geochemical survey (59 elements, 372 locations)” (Rodriguez-Galiano et al., 2015, p.804, p.816).</p> <p>“Gravity and magnetic survey (330 ground</p>

			stations)” (Rodriguez-Galiano et al., 2015, p.804, p.816).
<p>Davenport, T.H., 2014. What businesses can learn from sports analytics. MIT Sloan Management Review [online]. Available at:</p> <p><a href="https://sloanreview.mit.edu/article/what-businesses-can-learn-from-sports-analytics/">https://sloanreview.mit.edu/article/what-businesses-can-learn-from-sports-analytics/</a></p> <p>Puerzer, R.J., 2005. Baseball literary journals: The grass-roots literature of the game. NINE: A Journal of Baseball History and Culture, 14(1), pp.130-143. DOI: <a href="https://doi.org/10.1353/nin.2005.0050">https://doi.org/10.1353/nin.2005.0050</a></p> <p>Baumer, B., Zimbalist, A. (2014) Preface. The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball. Philadelphia: University of Pennsylvania Press, pp. ix-xiv. DOI: <a href="https://doi.org/10.9783/9780812209129.ix">https://doi.org/10.9783/9780812209129.ix</a></p> <p>St. John, A. (2013) Powered By Bill James And Friends, The Red Sox Win (Another) Moneyball World Series. Forbes [online]. Available at: <a href="https://www.forbes.com/sites/allenstjohn/2013/10/31/powerd-by-bill-james-and-friends-the-red-sox-win-another-money">https://www.forbes.com/sites/allenstjohn/2013/10/31/powerd-by-bill-james-and-friends-the-red-sox-win-another-money</a></p>	<p>These articles expand on Michael Lewis' book (2003) Moneyball by detailing how the evolutions of sports analytics emerged from fans, journalists, etc., from a “grassroots” level, and this was primarily how sports analytics emerged and was focused on the on-field metrics rather than the off-field and business metrics, which were only focused on after sports management adopted the use of data analytics in the 1990s and 2000s (Davenport, 2014, pp.10-13) (Puerzer, 2005, pp.130-134) (Baumer and Zimbalist, 2014, p.ix) (St. John, 2013, np.).</p>	N/A	N/A

<a href="#">ball-world-series/</a> [Accessed on: 23rd July 2024]			
Troilo, M., Bouchet, A., Urban, T.L. and Sutton, W.A. (2016) Perception, reality, and the adoption of business analytics: Evidence from North American professional sport organizations. <i>Omega</i> , 59, pp.72-83.	<p>This article focusses more on off-field analytics, such as how business analytics can improve organisations profits (Troilo et al., 2016, pp.72-75).</p> <p>Troilo et al. (2016, pp.72-75) explore and evaluate the use of business analytics in North American professional sports organisations, using empirical analysis to find out management's perceptions of the effects of business analytics and the financial effects since their organisations used these analytics.</p>	<p>The article surveyed managers and executives from different American sports franchises (Troilo et al., 2016, pp.72-75).</p>	<p>Surveys: 225 managers/executives from the National Football League, the National Basketball Association, etc. were surveyed (Troilo et al., 2016, pp.72-75).</p>
Holland, C.P., 2015, May. Internet and Social Media Strategy in Sports Marketing. In ECIS.	<p>Another use of off-field analytics is for sports marketing (Holland, 2015, pp.1-2, pp.13-15). Holland's (2015, pp.1-2, pp.13-15) article goes into depth on how sports federations and players can use social media to improve their marketing, as business analytics demonstrate that social media is having an increasing audience and revenue.</p>	<p>ComScore 2012, International Paralympic Press Release, Facebook, Twitter (cited in: Holland, 2015, pp.1-2, pp.13-15).</p>	<p>"The distribution of the top 100 sports websites worldwide," "Share of unique visitors for a sample of international sports federations 2007-2011," "The distribution of sports websites and their share of unique visitors," "Facebook popularity of major sports federations and individual athletes, July 2012," and "Twitter popularity of major sports federations and individual athletes, July 2012" (cited in: Holland, 2015, pp.5-14).</p>
7) Mead, J., O'Hare, A. and McMenemy, P., 2023. Expected goals in football: Improving model performance and demonstrating value. <i>Plos one</i> , 18(4), p.e0282295.	<p>Mead et al. (2023, pp.1-3, pp.16-18, pp.26-27) aim to improve the performance of the xG metric by using some unused variables, such as psychological effects.</p> <p>A criticism of sports analytics is its inability to determine outcomes due to "intangible," i.e., immeasurable qualities, which the article aims to address (Mead et al., 2023, pp.1-3, pp.16-18, pp.26-27).</p>	<p>Wyscout, FBref, Statsbomb, Transfermarkt (cited in: Mead et al., 2023, pp.1-12, pp.16-18, pp.26-29)</p>	<p>Elo Rating, Multi-Layer Perceptron (MLP), AdaBoost, XgBoost, PlayeRank, etc (Mead et al., 2023, pp.5-12, pp.16-18, pp.26-27).</p>

<p>5) Bartosz, Ć., Giełczyk, A. and Choraś, M., 2021. Who will score? a machine learning approach to supporting football team building and transfers. Entropy, 23(1), p.90. DOI: <a href="https://doi.org/10.3390/e23010090">https://doi.org/10.3390/e23010090</a></p>	<p>ML in football.</p> <p>The article examines how ML can be used for football, such as planning player transfers, and as a tool for professional scouts (Bartosz et al., 2021, p.1, pp.4-8).</p> <p>Defines a successful transfer as: High physical, psychological, and technical “aspects” and if the player’s aspects are higher than the team’s average aspects (Bartosz et al., 2021, p.1, pp.4-8).</p> <p>This is similar to the author’s player score, but the author will also use advanced statistics such as xG, as well as metrics not found in this paper, i.e., % of dribblers tackled (Bartosz et al., 2021, p.1, pp.4-8).</p>	<p>WhoScored, Transfermarkt, and Sofascore (cited in: Bartosz et al., 2021, p.1, pp.4-8).</p>	<p>1) Defined numerous parameters for player assessment. 2) Created three definitions of a successful player transfer (Bartosz et al., 2021, p.1, pp.4-8).</p> <p>Algorithms used: Random Forest, Naive Bayes, and AdaBoost to predict player transfer success (Bartosz et al., 2021, p.1, pp.4-8).</p> <p>The data used was realistic and publicly available to train and test the ML classifiers (Bartosz et al., 2021, p.1, pp.4-8).</p> <p>4700 players from 156 clubs belonging to the eight most popular leagues (based on UEFA ranking) (Bartosz et al., 2021, p.1, pp.4-8).</p>
<p>Thakkar, P. and Shah, M. (2021) An assessment of football through the lens of data science. Annals of Data Science, pp.1-14. DOI: 10.1007/s40745-021-00323-2</p>	<p>Thakkar and Shah’s (2021, pp.823-824, p.832) article covers the increasing use of data science in football, and how this could benefit the clubs, as it gives them a competitive advantage. They mention that include on field metric such as the player’s physiological data, but also the off field effects such as the socio-economic effects (Thakkar and Shah, 2021, pp.823-824, p.832).</p>	<p>N/A</p>	<p>N/A</p>
<p>Arkow, D., Jachuck, R., Kolar, L. (2023) MONEYBALL: AN ANALYSIS OF OVER AND UNDERRVALUED NBA STATISTICS. The Harvard College Sports Analysis Collective (HSAC) [online]. Available at: <a href="https://harvardsportsanalysis.o">https://harvardsportsanalysis.o</a></p>	<p>Similar to the authors’ data science project, the article by Arkow et al., (2023, np.) merges two datasets with one with on-field and one with off-field data to create one dataset to analyse how different basketball metric affects a player’s salary, thus displaying which metrics are potentially overvalued and undervalued by the basketball transfer market.</p>	<p>Basketball Reference, Spotrac (Arkow et al., 2023, np.).</p>	<p>Linear Regression (Arkow et al., 2023, np.).</p> <p>2021-2022 NBA season (Arkow et al., 2023, np.).</p>

<p><a href="https://www.semanticscholar.org/2023/05/moneyball-analysis-of-over-and-undervalued-nba-statistics/">rg/2023/05/moneyball-analysis-of-over-and-undervalued-nba-statistics/</a> [Accessed on: 15th July 2024]</p>			
---	--	--	--

## Appendix B Supplementary Literature

Srivastava et al. (2021) explain different sports analytics methods and how they can be applied in the sports industry. Srivastava et al. (2021, pp.1-2) give a breakdown of sports analytics, first by explaining the difference between on and off-field analytics, different ML techniques such as unsupervised ML, and clustering, regression, and classification. This paper's strength is the well-structured breakdown of data science concepts and application of these concepts through sports analytics with practical examples (Srivastava et al., 2021). A potential limitation of the paper is that, as a generalisation, it does not go into detail about any of the advanced statistics for specific sports; therefore, other papers in the literature will need to be consulted for further investigation and additional detail (Srivastava et al., 2021).

For a modern update on AI and its use in sports, particularly in player scouting, Mazari's (2024, pp.13-16) article provides in-depth detail on the potential of AI in sports, particularly in football, such as identifying unearthing talent, predicting potential superstars based on data, AI tactic recommendations, etc. However, an area of AI not examined by articles in the review so far is some of the ethical considerations, with Mazari emphasising the importance of ensuring privacy and security (Mazari, 2024, pp.15-17). Moreover, as AI becomes more complicated, it will be harder for laymen to understand AI, leading to transparency issues or the 'black box problem', with a potential solution being "Explainable AI" ("XAI"), techniques that explain how AI models come to their decisions, which will be necessary for making sure there is trust and fairness between AI experts and laymen (Mazari, 2024, p.17).

An interesting analysis of basketball data is Arkow et al.'s (2023, np.) article, as it expands the use of the Moneyball principle of using less traditional metrics to identify undervalued players in a team sport like basketball. The paper uses linear regression to measure the impact of the metric on salaries and VORP, e.g, if a 2 standard deviation increase in 'rebounds' results in a 2 standard deviation increase in VORP but only a 1 increase in a player's salary, that would mean the market undervalues the rebound metric when paying players (Arkow et al., 2023, np.). The paper found some interesting insights, such as 'turnovers' being one of the most undervalued metrics

(Arkow et al., 2023, np.). This demonstrates how algorithms (e.g, linear regression) can find patterns that a traditional scout might miss and how sports organisations can misallocate their salaries based on overvalued metrics (Arkow et al., 2023, np.).

A criticism of basketball analytics is that there is not enough emphasis on defensive metrics and that metrics are biased towards attacking players (Franks et al., 2015, pp.1-4). Franks et al.'s (2015, pp.1-2, pp.4-8) paper attempts to improve basketball analytics by creating new defensive metrics such as 'counterpoints' as a means to quantify defensive performances, while stating that future work is needed to improve defensive metrics by integrating defensive schemes and intent into the defensive measurements.

Thakkar and Shah (2021, p. 823) discuss how football clubs have embraced AI and ML to give them a competitive advantage over rivals.

Thakkar and Shah (2021, pp.825-526) do this by researching how statistics, data analysis, and ML techniques such as regression, etc., are used and how they have influenced decision-making, as well as the potential challenges. The study is also notable in mentioning the idea of "gatekeeping" in the sports analytics industry due to the rivalry in sports incentivizing the search for competitive advantages, in this case through AI and ML, which keeps the research and innovation by different football clubs kept secret due to the competitive nature of the industry, unlike an industry like healthcare, which is more collaborative (Thakkar and Shah, 2021, p.831). While Thakkar and Shah (2021, pp.827-830) cover the impact of data science on football and how ML methods can be applied to sports to provide insight and analysis, a potential gap is that they do not examine any advanced metrics or statistics.

Football has recently started to adopt some forms of composite statistics as examined by Bartosz et al. (2021, p.1, pp.4-8) and how ML can be used for football, such as planning player transfers, and as a tool for professional scouts. This is done by creating a feature, e.g, 'technical aspects', which are made up of variables such as goals scored, successful dribbles, etc., creating a composite statistic that sums up a player's ability (Bartosz et al., 2021, p.1, pp.4-8). This aspect, along with other aspects, was used to evaluate a player's transfer and whether it was successful or not, e.g, by comparing their aspects before and after the transfer (Bartosz et al., 2021, p.1, pp.4-8). This is similar to the author's 'player score', in addition to using advanced metrics such as xG, and defines the 'score' by positions (defender, attacker and midfielder) rather than the key aspects of a player's ability (technical, psychological and physical) (Bartosz et al., 2021, p.1, pp.4-8) (FBref, nd., np.). It must be mentioned that although Bartosz et al., (2021, p.1, pp.4-9) does use ML and composite statistics together, this is for classification purpose, while the author's project combines ML and composite statistics for regression purposes.

A criticism mentioned earlier was that sports analytics struggles to deal with intangible aspects of a player, aspects of the game that cannot be measured, such as psychological issues (Sarlis and Tjortjis, 2020, p.17).

Mead et al. (2023, pp.1-2) attempt to rectify this problem with sports analytics by improving the performance of the xG model by using some previously unused features, such as psychological effects. The “psychological” feature was made of variables such as “match attendance, match importance and goal differential,” with the paper concluding that the “psychological” feature was one of the most important variables for xG models, with further research potentially being able to expand the ability to integrate sports analytics with ‘intangible measures’ (Mead et al., 2023, p.8, pp.25-26).

### Appendix C Variable List

Variable	Data Type	Explanation
Aerial Duels Lost	Discrete	When a player unsuccessfully competes one vs one with another player to touch the ball (above elbow height) (Wyscout, nd., np.).
Aerial Duels Won	Discrete	When a player successfully competes one vs one with another player to touch the ball (above elbow height) (Wyscout, nd., np.).
Aerial Duels Won Percentage	Continuous	[Aerial Duels Won / (Aerial Duels won + Aerial Duels Lost)] * 100 (Wyscout, nd., np.).
Age	Continuous	The age of a player.
Assists	Discrete	The final pass from a player from the team prior to the goal being scored by the player’s teammate, or an “Own goal” by the opposition team (Wyscout, nd., np.).
Attack Field	Continuous	A composite statistics used by the author as a general indicator of a player's attacking ability (FBref, nd., np.). Different variables had different weighting by position (FBref, nd., np.). [goals + assists + penalty goals + pen_miss + Progressive Carries + Progressive Passes + Successful take ons + Successful take ons percentage + Penalties Won + Crosses + Foul won] (FBref, nd., np.).
Average Goal Kick Length	Continuous	The average length of a goal kick that a keeper makes over a season (measured in yards) (FBref, nd., np.).
Average Pass Length	Continuous	The average length of the pass that a keeper makes over a season (measured in yards) (FBref, nd., np.).
Ball Recoveries	Discrete	“Number of Loose Balls Recovered” (FBref, nd., np.).

Blocked Passes	Discrete	“Opposition passes that are stopped by a player standing in the way” (FBref, nd., np.)
Blocked Shots	Discrete	“Opposition shots that are stopped by a player standing in the way” (FBref, nd., np.)
Blocks	Discrete	“Number of times the ball is stopped by a player standing in the way” (FBref, nd., np.)
Carries	Discrete	“Number of times a player controls the ball with their feet” (FBref, nd., np.).
Carries into Final 1/3	Discrete	When a player “carries” or “dribbles” the ball into the opposition's final 1/3 of the pitch to get closer to the opposition's goal (FBref, nd., np.).
Carries into penalty box	<b>Discrete</b>	When a player “carries” or “dribbles” the ball into the opposition's penalty box to get closer to the opposition's goal (FBref, nd., np.).
Carries (Progressive)	Discrete	“Carries” or “dribbles” that move the ball to the oppositions' goal line by at least 10 yards or any carries into the penalty area (FBref, nd., np.).
Carries (Progressive Distance)	Continuous	Measured in yards, the total distance a player “carries” or “dribbles” the ball towards the opposition's goal (FBref, nd., np.).
Carries (Total Distance)	Continuous	Measured in yards, the total distance a player “carries” or “dribbles” the ball in any direction (FBref, nd., np.).
Challenges Lost	Discrete	This is the total “number of unsuccessful attempts to challenge a dribbling player” (FBref, nd., np.).
Clearance	Discrete	When a player tries to clear a ball away from their goal (Wyscout, nd., np.).
Crosses	Discrete	When the ball is played (“crossed”) from the offensive wide areas towards a teammate (usually a striker) in front of the opponent's goal (Wyscout, nd., np.).
Crosses Faced	Discrete	The total number of crosses that a goalkeeper has faced from opposition players over a season (FBref, nd., np.).
Crosses Stopped	Discrete	The number of crosses that a goalkeeper has stopped over a season, usually by catching or punching the football (FBref, nd., np.).
Crosses Stopped Percentage	Discrete	$[(\text{Crosses stopped}/\text{Crosses Faced}) * 100]$
Defence Field	Continuous	A composite statistics used by the author as a general indicator of a player's defensive ability (FBref, nd., np.). Different variables had different weighting by position (FBref, nd., np.). [Tackles_won + Interceptions + Blocks + Clearances + Challenges Lost + Challenges Attempted + Challenges Tkl percentage + Penalties Conceded + Own

		Goals + Ball Recoveries + Aerial Duels Won + Aerial Duels Lost + Aerial Duels Won Percentage] (FBref, nd., np.).
Dispossessed	Discrete	When a player who has the ball loses it after being tackled by an opposition player (ball being lost due to attempted take-ons does not count) (FBref, nd., np.).
Dribblers challenged	Discrete	[No. of unsuccessful challenges of dribblers + no. of successful challenges] (FBref, nd., np.).
Dribblers Tackled	Discrete	“Number of dribblers tackled” (FBref, nd., np.)
Errors	Discrete	“Mistakes” that led to an opposition taking a shot (FBref, nd., np.).
Expected Goals (npxG)	Continuous	Quantifies a chance’s quality by calculating the probability that it will be scored using data on similar shots (excluding penalty goals) (Opta, nd., np.).
Expected Goals (xG)	Continuous	Quantifies a chance’s quality by calculating the probability that it will be scored using data on similar shots (Opta, nd., np.).
Excepted Assist (xA)	Continuous	Quantifies the probability that a completed pass will become an “assist”, even if a shot doesn't happen (Opta, nd., np.).
Fouls Committed	Discrete	Fouls are when a player commits an unfair or unsportsmanlike play, which can include unnecessarily aggressive or dangerous contact, which could carry an increased risk of injury (Sapp, 2018, p.1346).
Fouls Won	Discrete	When a player gets fouled, resulting in their team winning a free kick, penalty, etc. (Sapp, 2018, p.1346).
Goal	Discrete	How a football game is won, whoever has the most “goal wins”. A goal is when the football crosses the goal line completely.
Goals and Assists	Discrete	G + A or “Goal Contributions = (Goals + Assist)
Goal Kicks Attempted	Discrete	The total number of kicks a goalkeeper makes after they have to restart play (FBref, nd., np.).
Goals minus PK	Discrete	Goals minus PK = (Goals - penalties)
Height in cm	Continuous	How tall a player is in CM.
Highest Market Value in EUR	Continuous	The highest expected value of a player in the football transfer market with its “free market” buying and selling players (Xerxes, nd., np.).
Interception	Discrete	When a player blocks a pass from an opposition player, winning possession (Wyscout, nd., np.).

Long ball Pass Completed Percentage	Continuous	A goalkeeper's "long ball" passing accuracy (FBref, nd., np.). A long ball is considered any pass by a keeper that is 40 or more yards (FBref, nd., np.). [(Long balls completed / All long balls) * 100] (FBref, nd., np.).
Market Value in Eur	Continuous	Or "estimated transfer value" is the expected value of a player in the football transfer market with its "free market" buying and selling players (Xerxes, nd., np.). "Both individual transfer modalities and situational conditions are relevant in determining market values" (Xerxes, nd., np.).
Miscontrols	Discrete	The amount of times a player fails to control a ball (FBref, nd., np.).
Number of Sweeps	Discrete	The number of defensive actions (e.g., tackles) a goalkeeper makes outside of their penalty area (FBref, nd., np.).
Number of Sweeps per 90	Continuous	The number of defensive actions (e.g., tackles) a goalkeeper makes outside of their penalty area every 90 minutes (FBref, nd., np.).
Offsides	Discrete	Amount of times a player is offside (FBref, nd., np.).
Off-the-line Field	Continuous	A composite statistics used by the author as a general indicator of a goalkeeper's "off the line" or "sweeping" ability (FBref, nd., np.) (Ghar et al., 2021, p.1163). Different variables had different weighting by position (FBref, nd., np.). [Cross stop + Sweep + long pass accuracy + touches] (FBref, nd., np.).
Own goals	Discrete	When a player scores a goal in their own goal net (usually by accident) (FBref, nd., np.).
Passes Attempted	Discrete	The number of passes that a player has attempted overall, regardless of distance (FBref, nd., np.).
Passes Receive	Discrete	The amount of times they received a pass from a teammate (FBref, nd., np.).
Penalty Kicks Attempted	Discrete	How many penalties attempts a goalkeeper faces (FBref, nd., np.).
Penalty Kicks Missed	Discrete	How penalties kick shots missed the goal entirely (Fbref, nd., np.).
Penalty Kick Save Percentage	Continuous	[(Penalties Saved by goalkeeper / Penalties attempted by opposition player)*100] (FBref, nd., np.).
Penalties Attempted	Discrete	The number of times a player has taken a penalty kick (FBref, nd., np.)
Penalties Conceded	Discrete	When a player gives away a penalty, usually by fouling a player in the penalty box (FBref, nd., np.).
Penalties Goals	Discrete	Goals that resulted from a penalty (FBref, nd., np.).
Penalties Saved	Discrete	The number of penalties that a goalkeeper has saved (FBref, nd., np.).

Performance Save Percentage	Continuous	$[(\text{Saves by Goalkeeper} / \text{Total Shots on target to goal faced}) * 100]$ (FBref, nd., np.).
Per 90 Min Gls	Continuous	Goals that have been scored by a player for every 90 minutes (FBref, nd., np.).
Per 90 Min Goal Contributions	Continuous	Goals and assists that have been scored by a player for every 90 minutes (FBref, nd., np.).
Per 90 Min Non Pen Goals	Continuous	Goals that have been scored by a player for every 90 minutes, excluding penalties (FBref, nd., np.).
Per 90 Min Non Penalty Goal Contributions	Continuous	Goals and assists that have been scored by a player for every 90 minutes, excluding penalties (FBref, nd., np.).
Per 90 Min npxG	Continuous	A player's non-penalty expected goals for every 90 minutes (FBref, nd., np.).
Per 90 Min npxG+xAG	Continuous	A player's non-penalty expected goals and expected assists for every 90 minutes (FBref, nd., np.).
Per 90 Min xAG	Continuous	A player's expected assists for every 90 minutes (FBref, nd., np.).
Per 90 Min xG	Continuous	A player's expected goals for every 90 minutes (FBref, nd., np.).
Per 90 Min xG+xAG	Continuous	A player's expected goals and expected assists for every 90 minutes (FBref, nd., np.).
Per 90 Minutes Ast	Continuous	A player's assists for every 90 minutes (FBref, nd., np.).
Per Dribblers Tkl Succ	Continuous	No. of dribblers tackled successfully / No. of dribblers tackled attempted (FBref, nd., np.).
Percentage of Goal Kicks Launched	Continuous	The percentage of goal kicks that were launched (40 or more yards) (FBref, nd., np.). $[(\text{Goal kicks launched} / \text{Goal kicks overall}) * 100]$ (FBref, nd., np.).
Percentage of Passes Launched	Continuous	The percentage of passes that were "long balls" or "launched," which is when a pass is 40 or more yards (FBref, nd., np.).
Performance CrdR	Discrete	A yellow card is a warning for reckless play, unsporting behaviour, or accumulation of fouls, (Sapp, 2018, p.1346).
Performance CrdY	Discrete	A red card is for more serious offences like dangerous challenges or violent conduct or an accumulation of two yellow cards (Sapp, 2018, p.1346).
Player Score	Continuous	A composite statistic that was created by the author for the project based on the Player Efficiency Rating metric found in basketball as a means to give a general performance level of a player over a season, relative to their

		peers in the same position (Hollinger, 2003, cited in: Seshadri, 2024, p.1-2). Player Score for four positions (Goalkeeper, Defender, Midfielder and Attacking Players).
Playing Time 90s	Continuous	How many minutes a player played for every 90 minutes (FBref, nd., np.).
Playing Time Min	Continuous	How many minutes a player played over a season (FBref, nd., np.).
Playing Time MP	Discrete	The number of matches played (FBref, nd., np.).
Playing Time Starts	Discrete	The number of matches they started (FBref, nd., np.).
Post-Shot-Expected-Goals Minus Goals Allowed	Continuous	This measures the difference between how many goals a keeper should have allowed and how many they actually allowed (FBref, nd., np.). If the number is positive, then they saved more goals than they should have (FBref, nd., np.). If it's negative, they allowed more goals to be scored than they should have (FBref, nd., np.).
Post-Shot-Expected-Goals Minus Goals Allowed Per 90	Continuous	This is a keeper's Post-Shot-Expected-Goals Minus Goals Allowed for every 90 mins they play (FBref, nd., np.).
Progression Progressive Passes	Discrete	Any pass into the penalty area or a pass that brings the ball "10 yards" closer to the goal line (FBref, nd., np.).
Progressive Passes Received	Discrete	A pass received by the player into the penalty area or a pass that brings the ball "10 yards" closer to the goal line (FBref, nd., np.).
Shot-Stopping Field	Continuous	A composite statistics used by the author as a general indicator of a goalkeeper's "shot-stopping" ability (FBref, nd., np.) (Ghar et al., 2021, p.1163). Different variables had different weighting by position (FBref, nd., np.). [Save percentage + Penalty Save Percentage + Number of penalty kicks saved + Post-Shot-Expected-Goals Minus Goals Allowed] (FBref, nd., np.)
Sweeper Average Distance	Continuous	The average distance of defensive actions that a goalkeeper makes away from their own goal (measured in yards).
Tackles Att 3rd	Discrete	Tackles made in the attacking third (opposition's side of the pitch) (FBref, nd., np.).
Tackles Def 3rd	Discrete	Tackles made in the defensive third (tackler's side of the pitch) (FBref, nd., np.).
Tackles Mid 3rd	Discrete	Tackles made in the middle third of the pitch (FBref, nd., np.).
Tackles Tkl	Discrete	The no. of tackles made by a player overall (FBref, nd., np.).
Take Ons Attempted	Discrete	Number of times a player attempted to "take-on" a player while dribbling, i.e., they tried to dribble past them (FBref, nd., np.).

Take Ons Successful	Discrete	Number of times a player successfully attempted to “take-on” a player while dribbling (FBref, nd., np.).
Take Ons Successful Percentage	Continuous	(Take Ons Att / Take Ons Succ) * 100 (FBref, nd., np.).
Take Ons Tkld	Discrete	The no. of times a player was tackled when attempting to take on an opposition player (FBref, nd., np.).
Take Ons Tkld Percentage	Continuous	(Take Ons Tkld / Take ons attempted) * 100 (FBref, nd., np.).
Tackles Plus Interceptions	Discrete	(Tackles + Interceptions) (FBref, nd., np.).
Touches	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Att 3rd	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” in the opposition’s attacking third of the pitch (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Att Pen	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” in the opposition’s penalty box (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Def 3rd	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” in the defensive area of the pitch (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Def Pen	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” in the team’s penalty box (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Live	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” but excludes touches that do not come from open play such as penalty kicks (Raumdata, 2020, np.) (FBref, nd., np.).
Touches Mid 3rd	Discrete	When a player “receives a pass, dribbles and then sends a pass or shot” in the middle third of the pitch (Raumdata, 2020, np.) (FBref, nd., np.).
Throws Attempted	Discrete	How throws a goalkeeper has attempted overall (FBref, nd., np.).
Usage Rate Per	Continuous	(Average touches per game by player / Average touches per game by the whole team) * 100 (Raumdata, 2020, np.). In other words, how much of the ball does a player have compared to the rest of the team (Raumdata, 2020, np.).

## Appendix D

### Player Score Criteria

```
def calculate_performance_attack(row):
    goals = (row['Goals_Minus_PK'] * 100)
    assists = row['Assists'] * 50
    pen_goals = row['Penalty_Goals'] * 50
    pen_miss = (row['Penalties_Attempted'] - row['Penalty_Goals']) * (-50)
    xG = row['Expected_xG'] * 2
    XG_GS = (row['Expected_xG'] - row['Goals']) * 5
    Prog_Carries = row['Carries_PrgC'] * 5
    Progression_PrgP = row['Progression_PrgP'] * 3
    Take_Ons_Succ = row['Take_Ons_Succ'] * 2
    Take_Ons_Succ_per = row['Take_Ons_Succ_Per'] * 10
    Tackles_won = row['Tackles_TklW']
    Interceptions = row['Interceptions']
    Blocks = row['Blocks']
    Clearances = row['Clearances']
    Crosses = row['Crosses']
    Foul_won = row['Fouls_Won']
    Offsides = row['Offsides'] * (-3)
    Penalties_Won = row['Penalties_Won'] * 5
    Penalties_Conceded = row['Penalties_Conceded'] * (-5)
    Own_Goals = row['Own_Goals'] * (-5)
    Ball_Recoveries = row['Ball_Recoveries']
    Aerial_Duels_Won = row['Aerial_Duels_Won']
    Aerial_Duels_Lost = row['Aerial_Duels_Lost']
    Aerial_Duels_Won_Per = row['Aerial_Duels_Won_Per'] * 10
    Challenges_Lost = row['Challenges_Lost'] * (-1)
    Y_card = row['Performance_CrdY'] * (-3)
    R_card = row['Performance_CrdR'] * (-5)
```

```

def calculate_performance_midfield(col_mid):
    goals = col_mid['Goals_Minus_PK'] * 50
    assists = col_mid['Assists'] * 50
    pen_goals = col_mid['Penalty_Goals'] * 30
    pen_miss = (col_mid['Penalties_Attempted'] - col_mid['Penalty_Goals']) * (-30)
    Prog_Carries = col_mid['Carries_PrgC'] * 10
    Progression_PrgP = col_mid['Progression_PrgP'] * 10
    Take_Ons_Succ = col_mid['Take_Ons_Succ'] * 2
    Take_Ons_Succ_per = col_mid['Take_Ons_Succ_Per'] * 10
    Tackles_won = col_mid['Tackles_TklW'] * 3
    Interceptions = col_mid['Interceptions'] * 3
    Blocks = col_mid['Blocks'] * 3
    Clearances = col_mid['Clearances'] * 3
    Crosses = col_mid['Crosses']
    Foul_won = col_mid['Fouls_Won']
    Offsides = col_mid['Offsides'] * (-3)
    Penalties_Won = col_mid['Penalties_Won'] * 5
    Penalties_Conceded = col_mid['Penalties_Conceded'] * (-5)
    Own_Goals = col_mid['Own_Goals'] * (-5)
    Ball_Recoveries = col_mid['Ball_Recoveries']
    Aerial_Duels_Won = col_mid['Aerial_Duels_Won']
    Aerial_Duels_Lost = col_mid['Aerial_Duels_Lost']
    Aerial_Duels_Won_Per = col_mid['Aerial_Duels_Won_Per'] * 10
    Challenges_Att = col_mid['Dribblers_Challenged'] * 10
    Challenges_Tkl_per = col_mid['Per_Dribblers_Tkl_Succ'] * 5
    Challenges_Lost = col_mid['Challenges_Lost'] * (-2)
    Y_card = col_mid['Performance_CrdY'] * (-3)
    R_card = col_mid['Performance_CrdR'] * (-5)

```

```

def calculate_performance_defence(col_def):
    goals_def = col_def['Goals_Minus_PK'] * 25
    assists_def = col_def['Assists'] * 25
    pen_goals_def = col_def['Penalty_Goals'] * 15
    pen_miss_def = (col_def['Penalties_Attempted'] - col_def['Penalty_Goals']) * (-15)
    Prog_Carries_def = col_def['Carries_PrgC'] * 3
    Progression_PrgP_def = col_def['Progression_PrgP'] * 3
    Take_Ons_Succ_def = col_def['Take_Ons_Succ']
    Take_Ons_Succ_per_def = col_def['Take_Ons_Succ_Per']
    Tackles_won_def = col_def['Tackles_TklW'] * 5
    Interceptions_def = col_def['Interceptions'] * 5
    Blocks_def = col_def['Blocks'] * 5
    Clearances_def = col_def['Clearances'] * 5
    Crosses = col_def['Crosses']
    Foul_won = col_def['Fouls_Won']
    Penalties_Won = col_def['Penalties_Won'] * 5
    Penalties_Conceded = col_def['Penalties_Conceded'] * (-5)
    Own_Goals = col_def['Own_Goals'] * (-5)
    Ball_Recoveries = col_def['Ball_Recoveries']
    Aerial_Duels_Won = col_def['Aerial_Duels_Won']
    Aerial_Duels_Lost = col_def['Aerial_Duels_Lost']
    Aerial_Duels_Won_Per = col_def['Aerial_Duels_Won_Per'] * 10
    Challenges_Att_def = (col_def['Dribblers_Challenged']) * 15
    Challenges_Tkl_per_def = (col_def['Per_Dribblers_Tkl_Succ']) * 25
    Challenges_Lost_def = col_def['Challenges_Lost'] * (-3)
    Y_card_def = col_def['Performance_CrdY'] * (-3)
    R_card_def = col_def['Performance_CrdR'] * (-5)

```

```

def calculate_performance_GK(col_GK):
    per_save = col_GK['Performance_Save_Per']
    pen_save_per = col_GK['Penalty_Kick_Save_Per'] * 5
    pen_kick_save = col_GK['Penalties_saved'] * 5
    Post_shot_ex = col_GK['PSXG_minus_GA'] * 10
    Cross_stop = col_GK['Crosses_Stp_Per']
    Sweep = col_GK['Number_of_Sweep'] * 5
    long_pass_acc = col_GK['Long_Pass_Acc_per'] * 5

    return per_save + pen_save_per + pen_kick_save + Post_shot_ex + Cross_stop + Sweep + long_pass_acc

```

## Appendix E

### Terminology

Term	Explanation
Centre Back/Central Defender	These defenders play central, usually tasked with dealing with the centre-forward (Ghar et al., 2021, p.1163).
Fullback	The defenders play on the left and right flanks and are usually tasked with dealing with wingers (Ghar et al., 2021, p.1163).
Inverted Fullback	These fullback instead of staying near the touch line or running up to join the the winger, they “invert” by joining the midfield and assisting the team by playing more centrally (Desmond, 2022, cited in: Plakias et al., 2023, p.9)
Midfielder	They play centrally in the pitch, in between the attackers and defenders, and usually have both defensive and attacking responsibilities (Ghar et al., 2021, p.1163).
Moneyball	A book originally written by Michael Lewis (2003) that details the baseball organisation, Oaklands Athletics, and their struggle to compete with the financially stronger baseball teams. The book explains how Oaklands Athletic was able to become one of the best teams in American baseball, through its General and Assistant General Manager, by using data analytics and statistical methods to gain a competitive advantage (Lewis, 2003). The term “Moneyball” is now used interchangeably with using analytics and statistics to gain an advantage in sports, especially over richer competitors (Lewis, 2003).
Player Score	A composite statistic that was created by the author for the project based on the Player Efficiency Rating metric found in basketball as a means to give a general performance level of a player over a season, relative to their peers in the same position (Hollinger, 2003, cited in: Seshadri, 2024, p.1-2). Player Score for four positions (Goalkeeper, Defender, Midfielder and Attacking Players).
Shot-Stopper	These keepers are primarily tasked with making saves and having good reflexes and are not expected to come off their line as much (Ghar et al., 2021, p.1163).
Striker/Centre Forward	Out of the attacking players they usually play the closest to the opposition’s goal and they are generally positioned centrally, their main task is normally to score goals (Ghar et al., 2021, p.1163).
Sweeper Keeper	These goalkeepers are effectively another defender, as they come off their line to tackle opposition players or catch crosses made by the opposition (Ghar et al., 2021, p.1163).

Transfer market	This is the football market where players are bought and sold for a fee, and where players who have expired contracts can go for free to the club that bids with the highest salaries (Depken and Globan, 2021, pp.889-890).
Transfermarkt	Not to be confused with the “Transfer market” (Smith, 2021, np.). A widely used and popular website which offers informations for tens of thousands players for different countries and leagues, and one of its most popular and some would suggest polarising metrics is their “estimated transfer value” or “market value” which is an estimate of how much a player is worth as determined by the website’s thousands of volunteers, carefully selected by the website’s staff (Smith, 2021, np.).
Wing back	These are fullbacks who usually play in line with their midfielders and play much more forward than a traditional fullback, giving them extra attacking responsibility. (Ghar et al., 2021, p.1163)
Winger/Wide forward	These are the attacking players that play on the flanks and have tasks such as dribbling and crossing the back to create attacking opportunities (Ghar et al., 2021, p.1163).

## **Appendix F. Ranking of Attacking Players’ features**

### **Ranking of Attacking Players’ features (Random Forest) on Player’s Predicted Value:**

1. Predicted\_Performance: 0.38210334811295027
2. Touches\_Att\_Pen: 0.10062245254197479
3. Score: 0.06516523502999304
4. Touches\_Att\_3rd: 0.04347742295927204
5. Carries\_Into\_Penalty\_Box: 0.04026964191468564
6. Age: 0.021479890674920496
7. Goals\_and\_Assists: 0.01949614236198304
8. Expected\_xAG: 0.013849608204622324
9. Progressive\_Passes\_Recieved: 0.01243146650446196
10. Per\_90\_Min\_npxG+xAG: 0.011792759059424056
11. Aerial\_Duels\_Won\_Per: 0.010563136798692845
12. Progression\_PrgP: 0.009716998059753158
13. Aerial\_Duels\_Lost: 0.009270553418151962
14. Per\_90\_Min\_xAG: 0.008424298501202155
15. Carries\_Into\_Final\_1/3: 0.008210292768905286
16. Take\_Ons\_Att: 0.007522415148181302
17. Goals: 0.007243362453157817
18. Playing\_Time\_MP: 0.00716100262279562
19. Dispossessed: 0.0070744425433726145
20. Progression\_PrgR: 0.0070405788568513525
21. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.006719497624750457

22. Take\_Ons\_Tkld\_Per: 0.006663332315547499  
23. Expected\_npxG: 0.0065076020759561185  
24. Take\_Ons\_Succ: 0.006417843631116958  
25. Fouls\_Committed: 0.006071434954258198  
26. Playing\_Time\_Starts: 0.005646075226395251  
27. Clearances: 0.005552526393107387  
28. Per 90 Minutes Ast: 0.005480047408785837  
29. Per\_90\_Min\_Non\_Pen\_Goals: 0.005263371497177089  
30. Offsides: 0.005229407594966093  
31. Usage\_Rate\_Per: 0.0051264102654732255  
32. Performance\_CrdY: 0.005080362480351138  
33. Take\_Ons\_Succ\_Per: 0.004836481697626587  
34. Goals\_Minus\_PK: 0.004799948098340275  
35. Interceptions: 0.0046321561587909675  
36. Carries: 0.004524616943114377  
37. Per\_90\_Min\_xG: 0.004486674534921929  
38. Height\_in\_CM: 0.004190510859960563  
39. Playing\_Time\_90s: 0.0040446451739959465  
40. Aerial\_Duels\_Won: 0.003951024207793445  
41. Carries\_PrgDist: 0.00394182589047037  
42. Passes\_Recieved: 0.003929589495442944  
43. Crosses: 0.00392078591075054  
44. Per\_90\_Min\_npxG: 0.003892184831585022  
45. Miscontrol: 0.0038629942654934742  
46. Take\_Ons\_Tkld: 0.0037615387985594382  
47. Tackles\_Mid\_3rd: 0.0037442596968864377  
48. Per\_90\_Min\_xG+xAG: 0.0037348841983775778  
49. Playing\_Time\_Min: 0.0035861670694466658  
50. Tackles\_Att\_3rd: 0.0033767056002491596  
51. Blocked\_Shots: 0.003348497314810913  
52. Per\_Dribblers\_Tkl\_Succ: 0.0033226573557634687  
53. Progression\_PrgC: 0.003161268847548694  
54. Per\_90\_Min\_Gls: 0.003117614571639607  
55. Tackles\_Def\_3rd: 0.003021682919597247  
56. Carries\_TotDist: 0.0029117227100495004  
57. Per\_90\_Min\_Goal\_Contributions: 0.0028946796284325406  
58. Touches\_Mid\_3rd: 0.002849048489361956  
59. Blocks: 0.002809896334676342  
60. Fouls\_Won: 0.002755481803323679  
61. Blocked\_Passes: 0.002743111648715116  
62. Carries\_PrgC: 0.0025081090950481873  
63. Touches\_Def\_3rd: 0.0024011388287478582  
64. Expected\_xG: 0.0023306809071806877  
65. Touches\_Def\_Pen: 0.0022700488276503226  
66. Ball\_Recoveries: 0.0021912214277608273  
67. Dribblers\_Challenged: 0.0018988213338780142  
68. Tackles\_TklW: 0.001867765682003564  
69. Assists: 0.0018635998553812785

70. Tackles\_Tkl: 0.0018610865031168265  
 71. Tkl\_Plus\_Int: 0.001765521139481408  
 72. Dribblers\_Tkl: 0.0017175469969021762  
 73. Challenges\_Lost: 0.0017059717532249097  
 74. Penalty\_Goals: 0.0015668712556277408  
 75. Touches: 0.0014027618408575325  
 76. Touches\_Live: 0.0010411379513889935  
 77. Errors: 0.0006511188668876852  
 78. Penalties\_Won: 0.0006184788735881563  
 79. Penalties\_Attempted: 0.0006019160048772225  
 80. Performance\_CrdR: 0.0005245670363966706  
 81. Penalties\_Conceded: 0.00038602076103806123  
 82. Own\_Goals: 0.0

#### **Ranking of Midfielders' features (Random Forest) on Player's Predicted Value:**

1. Touches\_Att\_3rd: 0.12476059363983796
2. Expected\_xAG: 0.05640201251723479
3. Touches\_Att\_Pen: 0.05619973977522036
4. Passes\_Received: 0.05368874101445893
5. Goals\_and\_Assists: 0.04808286592460368
6. Crosses: 0.03520568373602146
7. Carries\_Into\_Final\_1/3: 0.03109903199570955
8. Tackles\_Att\_3rd: 0.030773996305091852
9. Carries\_PrgDist: 0.030520950082270838
10. Age: 0.03034758770860742
11. Carries\_Into\_Penalty\_Box: 0.024886302240488146
12. Expected\_xG: 0.01909069690609688
13. Aerial\_Duels\_Won\_Per: 0.01818949734857484
14. Predicted\_Performance: 0.017637155873366468
15. Fouls\_Won: 0.016421215261066463
16. Usage\_Rate\_Per: 0.015800560813427005
17. Aerial\_Duels\_Lost: 0.015217682911566697
18. Touches\_Def\_Pen: 0.0128391719013872
19. Score: 0.01265521524874483
20. Playing\_Time\_MP: 0.012601963832943586
21. Carries: 0.01177173639092957
22. Height\_in\_CM: 0.010915010801260418
23. Per\_90\_Min\_Gls: 0.01073159862030896
24. Dispossessed: 0.010160587720623115
25. Per\_90\_Min\_xG+xAG: 0.009739312445943834
26. Challenges\_Lost: 0.009678931580253422
27. Assists: 0.00950482997208067
28. Take\_Ons\_Att: 0.008700399814653051
29. Tackles\_Mid\_3rd: 0.00808842380391353
30. Offsides: 0.00801034697048481
31. Per\_90\_Min\_Non\_Pen\_Goals: 0.007897254729716634
32. Per\_90\_Min\_npxG+xAG: 0.007315350035296638

33. Per\_90\_Min\_npxG: 0.007125756699801004  
34. Ball\_Recoveries: 0.007116893615626843  
35. Fouls\_Committed: 0.007076411453568948  
36. Take\_Ons\_Tkld\_Per: 0.006841136726969031  
37. Take\_Ons\_Succ\_Per: 0.006819077795170326  
38. Miscontrol: 0.006742935228320332  
39. Dribblers\_Challenged: 0.006697347132137684  
40. Performance\_CrdY: 0.006551805462677079  
41. Per\_Dribblers\_Tkl\_Succ: 0.00647283793545886  
42. Progression\_PrgC: 0.006323198235914399  
43. Tackles\_Def\_3rd: 0.006221759654320146  
44. Carries\_TotDist: 0.0061666433993923745  
45. Expected\_npxG: 0.006086681507077664  
46. Progression\_PrgP: 0.005781706820017397  
47. Take\_Ons\_Tkld: 0.005728787842793692  
48. Progressive\_Passes\_Recieved: 0.005461402151081717  
49. Per\_90\_Min\_xAG: 0.005341823598459246  
50. Dribblers\_Tkl: 0.00528479588572213  
51. Take\_Ons\_Succ: 0.0051331967501080395  
52. Tkl\_Plus\_Int: 0.0050852202351181885  
53. Blocked\_Shots: 0.005021084475880914  
54. Carries\_PrgC: 0.005007628730106688  
55. Tackles\_Tkl: 0.004967798833594738  
56. Interceptions: 0.00488963248490527  
57. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.004844147566264395  
58. Playing\_Time\_90s: 0.004824751043814587  
59. Playing\_Time\_Min: 0.004820994301993695  
60. Per\_90\_Min\_xG: 0.004638899539705425  
61. Tackles\_TklW: 0.004583895147294711  
62. Penalties\_Won: 0.0043181132164056245  
63. Touches\_Def\_3rd: 0.004315913715305906  
64. Touches: 0.004120373793612618  
65. Aerial\_Duels\_Won: 0.003960536613598939  
66. Blocked\_Passes: 0.0038785977207148894  
67. Goals\_Minus\_PK: 0.003874884032165772  
68. Touches\_Live: 0.003754872570552861  
69. Touches\_Mid\_3rd: 0.0036174378429609778  
70. Per\_90\_Min\_Goal\_Contributions: 0.00350850847949299  
71. Errors: 0.0034486649099676205  
72. Playing\_Time\_Starts: 0.0027512788007629575  
73. Blocks: 0.0026907571276031717  
74. Per 90 Minutes Ast: 0.002556562621377428  
75. Clearances: 0.0024877807256053538  
76. Progression\_PrgR: 0.0024592053372280392  
77. Goals: 0.0022373892114762235  
78. Performance\_CrdR: 0.0017619576686524264  
79. Penalty\_Goals: 0.0005974846413872576  
80. Penalties\_Conceded: 0.0005865810020736104

81. Penalties\_Attempted: 0.00040119712348242844
82. Own\_Goals: 7.920467009363624e-05

**Ranking of Defenders' features (Random Forest) on Player's Predicted Value:**

1. Passes\_Received: 0.26352312553819524
2. Age: 0.08170985851305104
3. Carries: 0.03850027805555681
4. Touches\_Mid\_3rd: 0.025320090278763488
5. Usage\_Rate\_Per: 0.024818882431012813
6. Per\_90\_Min\_xAG: 0.023687220992517754
7. Fouls\_Committed: 0.01710489168447925
8. Clearances: 0.01684627253831757
9. Carries\_PrgDist: 0.01630973614855288
10. Tackles\_Def\_3rd: 0.016075393661311273
11. Blocks: 0.015687402166765778
12. Per\_Dribblers\_Tkl\_Succ: 0.015675846771462644
13. Aerial\_Duels\_Won: 0.013583890571980077
14. Aerial\_Duels\_Won\_Per: 0.013027550519644649
15. Height\_in\_CM: 0.012835444131926435
16. Playing\_Time\_MP: 0.012605453574406528
17. Per\_90\_Min\_npxG+xAG: 0.012152746021043024
18. Score: 0.011960397974514264
19. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.011373298778178286
20. Interceptions: 0.010996741653339817
21. Aerial\_Duels\_Lost: 0.01052712725718563
22. Fouls\_Won: 0.01045725165173128
23. Blocked\_Shots: 0.010039049545411296
24. Dribblers\_Challenged: 0.0097198290816324
25. Predicted\_Performance: 0.00966111800858837
26. Touches\_Att\_Pen: 0.00960612098229774
27. Playing\_Time\_90s: 0.009560940910175186
28. Tkl\_Plus\_Int: 0.00953557196689586
29. Per\_90\_Min\_xG: 0.00936070799078425
30. Take\_Ons\_Tkld\_Per: 0.009225807449714704
31. Playing\_Time\_Min: 0.008555083616025649
32. Expected\_xAG: 0.008552751602450113
33. Per\_90\_Min\_npxG: 0.008234124092960687
34. Carries\_Into\_Final\_1/3: 0.007982559461428574
35. Crosses: 0.007979962463555089
36. Playing\_Time\_Starts: 0.007914688443718896
37. Progression\_PrgP: 0.007905478501953684
38. Tackles\_Mid\_3rd: 0.007270173954585659
39. Dispossessed: 0.007241365051579075
40. Carries\_PrgC: 0.007221748494204915
41. Offsides: 0.007121764536537133
42. Touches\_Att\_3rd: 0.007096352825763309

43. Take\_Ons\_Succ\_Per: 0.006884579121263439  
 44. Per\_90\_Min\_Goal\_Contributions: 0.006833217323196972  
 45. Touches\_Def\_Pen: 0.006738882398929003  
 46. Blocked\_Passes: 0.006442225762304989  
 47. Touches\_Live: 0.006066167735267196  
 48. Challenges\_Lost: 0.006041368574353168  
 49. Per\_90\_Min\_xG+xAG: 0.00584211189174691  
 50. Tackles\_Tkl: 0.005680957368569241  
 51. Progression\_PrgC: 0.0055363981817093426  
 52. Performance\_CrdY: 0.005479885134369536  
 53. Carries\_TotDist: 0.005460666104468726  
 54. Ball\_Recoveries: 0.0053711833451939904  
 55. Expected\_npxG: 0.00532034932766631  
 56. Dribblers\_Tkl: 0.005264508690839477  
 57. Per 90 Minutes Ast: 0.005021204691989931  
 58. Carries\_Into\_Penalty\_Box: 0.004951347513419434  
 59. Expected\_xG: 0.004625633513443112  
 60. Touches: 0.004559149059734235  
 61. Miscontrol: 0.004533036271376937  
 62. Take\_Ons\_Tkld: 0.004485887258769346  
 63. Touches\_Def\_3rd: 0.004367942093842559  
 64. Tackles\_TklW: 0.004239708160647353  
 65. Progressive\_Passes\_Recieved: 0.0038822007777256466  
 66. Take\_Ons\_Att: 0.0037032409305693807  
 67. Per\_90\_Min\_Non\_Pen\_Goals: 0.0036824701742175616  
 68. Take\_Ons\_Succ: 0.003139303753591168  
 69. Per\_90\_Min\_Gls: 0.0030214458707318857  
 70. Tackles\_Att\_3rd: 0.002682571188553003  
 71. Errors: 0.0025742380468362523  
 72. Own\_Goals: 0.002392751596452691  
 73. Goals\_and\_Assists: 0.0019463528951637013  
 74. Progression\_PrgR: 0.0019458971293509811  
 75. Assists: 0.0019029581015213202  
 76. Performance\_CrdR: 0.0013058058953928076  
 77. Penalties\_Conceded: 0.001186016473177805  
 78. Goals\_Minus\_PK: 0.0010096376124586897  
 79. Goals: 0.0006661500816737572  
 80. Penalties\_Won: 0.0006429095198395991  
 81. Penalties\_Attempted: 5.3089900964192576e-06  
 82. Penalty\_Goals: 2.315453470472513e-07

#### **Ranking of Goalkeepers' features (Random Forest) on Player's Predicted Value:**

1. PSxG\_minus\_GA: 0.14945607016710385
2. Avg\_Pass\_Length: 0.08332891617006677
3. Throws\_Attemp: 0.08086606957933076
4. Crosses\_Stp: 0.07712300568660688

5. Per\_Passes\_Launched: 0.06612566827370162
6. Crosses\_Stp\_Per: 0.0649613194881085
7. Score: 0.05699796647239749
8. Crosses\_Faced: 0.05666166048608115
9. Number\_of\_Sweep: 0.0449361224704469
10. Long\_Pass\_Acc\_per: 0.03592120300713704
11. Passes\_Attempted: 0.02927287991572892
12. Usage\_Rate\_Per: 0.028567908272399918
13. Rk: 0.025240026559208522
14. Predicted\_Performance: 0.025126224263499624
15. Height\_in\_CM: 0.023743569957061065
16. Age: 0.02172888735149777
17. Touches: 0.02132011344644986
18. Avg\_Length\_Goal\_Kick: 0.0178643476049122
19. PSxG\_minus\_GA\_90: 0.01752881536640791
20. Goal\_Kicks\_Att: 0.016332313752851813
21. Sweeper\_Avg\_Dist: 0.014167339212427615
22. Performance\_Save\_Per: 0.013559479037300277
23. Number\_of\_Sweep\_per\_90: 0.01130944225463457
24. Penalty\_Kick\_Save\_Per: 0.008842287362224288
25. Per\_Goal\_Kicks\_Launch: 0.004446210598439507
26. Penalties\_saved: 0.0037772135743571783
27. Penalty\_Kicks\_PKm: 0.0007949396696180161

#### **Ranking of Attacking Players features (Random Forest) on Player's Predicted Performance:**

1. Goals\_and\_Assists: 0.6001345702845441
2. Playing\_Time\_Min: 0.07567456384502377
3. Touches\_Att\_3rd: 0.06466311924670279
4. Playing\_Time\_90s: 0.048863788074258944
5. Passes\_Received: 0.0455408565745604
6. Take\_Ons\_Succ\_Per: 0.02356950352120709
7. Aerial\_Duels\_Won\_Per: 0.018501389512019704
8. Carries: 0.011608586577489595
9. Carries\_TotDist: 0.008343215670953839
10. Aerial\_Duels\_Won: 0.007612698423414088
11. Take\_Ons\_Succ: 0.007252175973003725
12. Touches\_Mid\_3rd: 0.006466666525269522
13. Goals\_Minus\_PK: 0.005480336926895388
14. Usage\_Rate\_Per: 0.005233875688606616
15. Touches\_Att\_Pen: 0.005206333147730201
16. Carries\_PrgDist: 0.004701305632159927
17. Touches: 0.0036862752082433875
18. Touches\_Live: 0.0030625471459120547
19. Progression\_PrgR: 0.003049499961459485
20. Carries\_Into\_Final\_1/3: 0.0026096122888182825
21. Miscontrol: 0.0026045246883670235

22. Market\_Value\_in\_Eur: 0.002456825543028697  
23. Goals: 0.0023954677671293926  
24. Take\_Ons\_Tkld\_Per: 0.002238306105272983  
25. Highest\_Market\_Value\_in\_Eur: 0.0021466660302228747  
26. Per\_90\_Min\_Goal\_Contributions: 0.0016646570570350374  
27. Playing\_Time\_Starts: 0.0014552445269636588  
28. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.0014367001221085162  
29. Ball\_Recoveries: 0.0014256076617425691  
30. Fouls\_Committed: 0.0014099261377627365  
31. Touches\_Def\_3rd: 0.0013182606136274162  
32. Carries\_Into\_Penalty\_Box: 0.0012971454139903069  
33. Dispossessed: 0.0012775324878271787  
34. Per\_90\_Min\_Non\_Pen\_Goals: 0.0012531365711973908  
35. Fouls\_Won: 0.001093761101461027  
36. Expected\_xG: 0.001083476195419452  
37. Expected\_xAG: 0.0010751866705537966  
38. Per\_90\_Min\_Gls: 0.0009907148932534271  
39. Per\_90\_Min\_npxG+xAG: 0.0009902045794799802  
40. Progression\_PrgP: 0.0009353965758915729  
41. Expected\_npxG: 0.0007865523294910019  
42. Age: 0.0007791248582238586  
43. Clearances: 0.0007718275015833384  
44. Per\_Dribblers\_Tkl\_Succ: 0.0007668473511571318  
45. Aerial\_Duels\_Lost: 0.0007663911933514092  
46. Take\_Ons\_Att: 0.0007130023794531383  
47. Progression\_PrgC: 0.0007079330825807756  
48. Tackles\_Tkl: 0.0006564829257639218  
49. Take\_Ons\_Tkld: 0.0006444602702708552  
50. Carries\_PrgC: 0.0006431751037908991  
51. Progressive\_Passes\_Recieved: 0.0006268094142327766  
52. Touches\_Def\_Pen: 0.0005700172304783707  
53. Per\_90\_Min\_npxG: 0.0005533805949681817  
54. Playing\_Time\_MP: 0.0005352443555805505  
55. Tkl\_Plus\_Int: 0.0005228922618827659  
56. Per\_90\_Min\_xG+xAG: 0.0005029049687703332  
57. Tackles\_Mid\_3rd: 0.0004724880050134271  
58. Height\_in\_CM: 0.00045980446397734154  
59. Crosses: 0.00045907559346960243  
60. Blocked\_Passes: 0.00044027894999823785  
61. Blocks: 0.0004352496317434787  
62. Tackles\_Def\_3rd: 0.00043314176539983504  
63. Per\_90\_Min\_xG: 0.0004256962233064054  
64. Offsides: 0.00042367439606436823  
65. Tackles\_Att\_3rd: 0.0004150921655614563  
66. Interceptions: 0.0003948404848896079  
67. Tackles\_TklW: 0.0003854176771580615  
68. Challenges\_Lost: 0.000374789070938876  
69. Per\_90\_Min\_xAG: 0.0003650555421934985

70. Per 90 Minutes Ast: 0.00035294637345370766
71. Dribblers\_Challenged: 0.00034837733654914183
72. Dribblers\_Tkl: 0.0003125872371795559
73. Performance\_CrdY: 0.0002703949089673739
74. Blocked\_Shots: 0.00023976576319622022
75. Assists: 0.00022438933875438408
76. Penalties\_Won: 0.00012596800241200984
77. Penalty\_Goals: 0.00010223525998177252
78. Penalties\_Attempted: 8.299507573506737e-05
79. Performance\_CrdR: 4.192615994486428e-05
80. Errors: 3.186202538321511e-05
81. Penalties\_Conceded: 2.4037023563028736e-05
82. Own\_Goals: 1.2047329782069239e-06

#### **Ranking of Midfielders features (Random Forest) on Player's Predicted Performance:**

1. Touches: 0.31598405475935076
2. Touches\_Live: 0.31482331002044983
3. Progression\_PrgP: 0.16156663127968138
4. Usage\_Rate\_Per: 0.041971448711542465
5. Playing\_Time\_Min: 0.02577894497020885
6. Carries: 0.022798299954626382
7. Aerial\_Duels\_Won\_Per: 0.015235970787717979
8. Playing\_Time\_90s: 0.008534847007852085
9. Tackles\_Tkl: 0.007973105668828827
10. Tkl\_Plus\_Int: 0.006868490581115291
11. Take\_Ons\_Succ\_Per: 0.006723769189048771
12. Carries\_PrgDist: 0.005911413432003317
13. Ball\_Recoveries: 0.004047576311323254
14. Touches\_Att\_3rd: 0.0035329567113471564
15. Take\_Ons\_Succ: 0.0029681598048899773
16. Touches\_Mid\_3rd: 0.0027450689452123097
17. Tackles\_TklW: 0.0026255537785092138
18. Aerial\_Duels\_Won: 0.0025980729019153964
19. Dispossessed: 0.0021815262995446426
20. Per\_Dribblers\_Tkl\_Succ: 0.001882488602247513
21. Passes\_Recieved: 0.0018556051488155952
22. Dribblers\_Tkl: 0.0016615971016493462
23. Expected\_xAG: 0.001535224433303668
24. Playing\_Time\_Starts: 0.001508130008764713
25. Tackles\_Def\_3rd: 0.0014230926999946332
26. Highest\_Market\_Value\_in\_Eur: 0.0014223214974779445
27. Goals\_and\_Assists: 0.001322725417699484
28. Carries\_Into\_Final\_1/3: 0.0013160176993421366
29. Touches\_Att\_Pen: 0.0012734597458534179

30. Expected\_npxG: 0.0012477220701839755  
31. Take\_Ons\_Tkld\_Per: 0.0012291648823762108  
32. Goals: 0.0011658619847462652  
33. Age: 0.0010106199539787078  
34. Per\_90\_Min\_xG+xAG: 0.0009623391387074656  
35. Crosses: 0.000955869436862467  
36. Per\_90\_Min\_xAG: 0.0009532348176009885  
37. Height\_in\_CM: 0.0009353052084487327  
38. Dribblers\_Challenged: 0.0008699239990047753  
39. Goals\_Minus\_PK: 0.0008438529554460466  
40. Interceptions: 0.0008428522553703809  
41. Per\_90\_Min\_Gls: 0.0008413988827633308  
42. Carries\_TotDist: 0.0008306580529541885  
43. Miscontrol: 0.0008221127140085673  
44. Assists: 0.0008096184273288929  
45. Challenges\_Lost: 0.000716690736518376  
46. Market\_Value\_in\_Eur: 0.0006844880888288105  
47. Tackles\_Att\_3rd: 0.0006710101377306289  
48. Fouls\_Committed: 0.0006663802295852439  
49. Progression\_PrgC: 0.000663635764531625  
50. Take\_Ons\_Tkld: 0.0006450257802018057  
51. Blocks: 0.0006441206526799708  
52. Carries\_Into\_Penalty\_Box: 0.0006138189348175534  
53. Fouls\_Won: 0.0005698997739137645  
54. Per\_90\_Min\_Goal\_Contributions: 0.0005635806132742805  
55. Progressive\_Passes\_Recieved: 0.0005470762648307306  
56. Per\_90\_Min\_npxG+xAG: 0.0005460335584460441  
57. Playing\_Time\_MP: 0.0005198148759389399  
58. Expected\_xG: 0.0005145466402960753  
59. Per\_90\_Min\_npxG: 0.0005081355902157057  
60. Take\_Ons\_Att: 0.0005007829978068809  
61. Blocked\_Shots: 0.0004911134812748712  
62. Touches\_Def\_3rd: 0.0004676344113658654  
63. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.0004531018795545228  
64. Blocked\_Passes: 0.00042661789103620463  
65. Tackles\_Mid\_3rd: 0.00041518556754537587  
66. Carries\_PrgC: 0.00041146656301864986  
67. Aerial\_Duels\_Lost: 0.0003952469382282821  
68. Progression\_PrgR: 0.00039420706052467207  
69. Touches\_Def\_Pen: 0.00036133812984372914  
70. Clearances: 0.0003243535304844995  
71. Per\_90\_Min\_xG: 0.00029233904965161503  
72. Per 90 Minutes Ast: 0.0002734873825525433  
73. Offsides: 0.0002623975875709948  
74. Per\_90\_Min\_Non\_Pen\_Goals: 0.00023387309527556465  
75. Performance\_CrdY: 0.0002335806246474415  
76. Penalties\_Conceded: 0.000169411406610923  
77. Penalties\_Attempted: 0.00015322050894352475

78. Errors: 0.00010211778527566081  
79. Penalties\_Won: 9.094831635080934e-05  
80. Performance\_CrdR: 2.8818177502294392e-05  
81. Penalty\_Goals: 2.7080304424828134e-05  
82. Own\_Goals: 2.7021418576557964e-05

#### **Ranking of Defender features (Random Forest) on Player's Predicted Performance:**

1. Tkl\_Plus\_Int: 0.28058971549980155
2. Usage\_Rate\_Per: 0.27508658604555297
3. Per\_Dribblers\_Tkl\_Succ: 0.16686992672997877
4. Playing\_Time\_Min: 0.08392932688419075
5. Ball\_Recoveries: 0.03756121345525728
6. Playing\_Time\_90s: 0.03073942562955634
7. Dribblers\_Tkl: 0.022865465911195867
8. Aerial\_Duels\_Won\_Per: 0.011041180567038553
9. Aerial\_Duels\_Won: 0.010494717476978497
10. Touches\_Mid\_3rd: 0.007807347002230863
11. Tackles\_Def\_3rd: 0.007270889999407923
12. Clearances: 0.005711491400641255
13. Touches\_Live: 0.003479673335488754
14. Tackles\_TklW: 0.00330894851597299
15. Touches\_Def\_Pen: 0.0032709898167545466
16. Dribblers\_Challenged: 0.0030185467139819816
17. Touches: 0.0029596243439729756
18. Touches\_Def\_3rd: 0.0028842794630046546
19. Tackles\_Tkl: 0.00250155422172319
20. Touches\_Att\_Pen: 0.0021433545382432305
21. Blocks: 0.002054505718032207
22. Interceptions: 0.0018127758672947234
23. Blocked\_Shots: 0.0017344775029386166
24. Playing\_Time\_MP: 0.0016767639380936253
25. Passes\_Recieved: 0.0013576569985546967
26. Tackles\_Mid\_3rd: 0.0013377117031479146
27. Progression\_PrgP: 0.001262956042752088
28. Playing\_Time\_Starts: 0.001252063189263445
29. Carries\_TotDist: 0.001203190973216659
30. Market\_Value\_in\_Eur: 0.0010698988801741583
31. Take\_Ons\_Succ\_Per: 0.0009177851582696373
32. Highest\_Market\_Value\_in\_Eur: 0.0009159892065734351
33. Height\_in\_CM: 0.0008874840304237154
34. Aerial\_Duels\_Lost: 0.0008844343975613031
35. Blocked\_Passes: 0.0008642887479935998
36. Carries: 0.000861178956722134
37. Fouls\_Won: 0.0008297577953156597

38. Challenges\_Lost: 0.0007842984458328442  
39. Carries\_PrgDist: 0.000775993393877684  
40. Take\_Ons\_Tkld\_Per: 0.0007665531356417267  
41. Age: 0.0007641518974917575  
42. Expected\_xG: 0.0006906402720182173  
43. Fouls\_Committed: 0.0006515958405753286  
44. Touches\_Att\_3rd: 0.0006178889883691627  
45. Expected\_npxG: 0.0005715017053352388  
46. Dispossessed: 0.000558166188139759  
47. Take\_Ons\_Att: 0.0005480786612091928  
48. Miscontrol: 0.000508473296008343  
49. Crosses: 0.00041209342778051336  
50. Performance\_CrdY: 0.00040740758107539243  
51. Per\_90\_Min\_xAG: 0.0003716132501004774  
52. Take\_Ons\_Tkld: 0.00037005661961557923  
53. Per\_90\_Min\_npxG+xAG: 0.00035967690395780495  
54. Carries\_PrgC: 0.0003502779453301538  
55. Per\_90\_Min\_Goal\_Contributions: 0.00034693549936057355  
56. Per\_90\_Min\_xG: 0.0003391668395473764  
57. Progression\_PrgC: 0.00033848163458747194  
58. Carries\_Into\_Final\_1/3: 0.00032614674020973707  
59. Offsides: 0.0003164454030986398  
60. Per\_90\_Min\_xG+xAG: 0.0002981249307702078  
61. Carries\_Into\_Penalty\_Box: 0.00028808782872893223  
62. Take\_Ons\_Succ: 0.0002851476672553377  
63. Per 90 Minutes Ast: 0.00028389116542302904  
64. Per\_90\_Min\_npxG: 0.000282073745871053  
65. Expected\_xAG: 0.0002809865440301  
66. Penalties\_Conceded: 0.0002537399897197178  
67. Per\_90\_Min\_Non\_Penalty\_Goal\_Contributions: 0.00024705818645018903  
68. Progressive\_Passes\_Recieved: 0.00023420314476448007  
69. Tackles\_Att\_3rd: 0.0002289916157698511  
70. Errors: 0.00022591529859661706  
71. Progression\_PrgR: 0.00022380481834113813  
72. Goals\_and\_Assists: 0.0002111528228779611  
73. Assists: 0.00019038919013351395  
74. Per\_90\_Min\_Non\_Pen\_Goals: 0.0001733004705788517  
75. Goals: 0.0001632807157417179  
76. Per\_90\_Min\_Gls: 0.00015087055410680604  
77. Goals\_Minus\_PK: 0.00013331481478841677  
78. Performance\_CrdR: 7.123475051948014e-05  
79. Penalties\_Won: 6.665339713938109e-05  
80. Own\_Goals: 6.578730143359985e-05  
81. Penalties\_Attempted: 9.11778070887258e-06  
82. Penalty\_Goals: 5.2939787178581853e-08

**Ranking of Goalkeeper features (Random Forest) on Player's Predicted Performance:**

1. Number\_of\_Sweep: 0.34022443121906387
2. Penalty\_Kick\_Save\_Per: 0.17559121517978862
3. Passes\_Attempted: 0.17081145697644193
4. Long\_Pass\_Acc\_per: 0.09726951964532328
5. Penalties\_saved: 0.04111989300719873
6. PSxG\_minus\_GA: 0.03154113828210581
7. Number\_of\_Sweep\_per\_90: 0.020029200001841167
8. Touches: 0.019917562206199467
9. Sweeper\_Avg\_Dist: 0.015443659872627611
10. Per\_Goal\_Kicks\_Launch: 0.012865660297206462
11. PSxG\_minus\_GA\_90: 0.011006547278795993
12. Performance\_Save\_Per: 0.010489966387244592
13. Rk: 0.009156819608580455
14. Usage\_Rate\_Per: 0.005742735574481902
15. Per\_Passes\_Launched: 0.0056305063382546434
16. Crosses\_Stp: 0.004621546872312889
17. Avg\_Pass\_Length: 0.004176407622349701
18. Throws\_Attemp: 0.003747313105300677
19. Market\_Value\_in\_Eur: 0.0031038106610835887
20. Avg\_Length\_Goal\_Kick: 0.003037874398211936
21. Height\_in\_CM: 0.0028771335712759155
22. Crosses\_Stp\_Per: 0.0025090679606711496
23. Age: 0.002273181871766374
24. Highest\_Market\_Value\_in\_Eur: 0.0022184760310250567
25. Crosses\_Faced: 0.0021532811215351742
26. Goal\_Kicks\_Att: 0.0019418969285787658
27. Penalty Kicks PKm: 0.000499697980734263

## **Appendix G managerial applications the usage rate percentage and the moneyball map.**

### **Appendix G. Intro**

The three supplementary applications are:

1. The Attack/Defence 4 quadrant grid
2. ‘Usage Rate Percentage’ and its’ impact on a player’s performance (Raumdata, 2020, np.).
3. Create a visualisation map of the nationalities with the highest quantity of talented players. Categorises the countries with the best talent into a “Moneyball Map” (Lewis, 2003).

The first supplementary application is an Attack/Defence Grid for each position, in which players will be plotted for each position, e.g, for players in the attacking position. It will provide four quadrants: 1) attacking players who contribute to the team by attacking and defending; 2) attacking players who contribute solely to attacking while lacking in defensive contributions; 3) attacking players who contribute by defending but are lacking in attack; 4) attacking players who are lacking in both attack and defence.

Another useful managerial application is the “Usage Rate Percentage” (USP) impact on a player’s performance, a basketball metric adapted to football using the formula provided by Raumdata (2020, np.). Over a season, a player’s “USP” is the percentage of a team’s average total touches on the ball over a season that an individual player is responsible for, with a touch being defined as a player “receiving a pass, dribbling and then sending a pass or shot” (Raumdata, 2020, np.). Therefore, for the project, it would be interesting to observe a possible relationship between performance and the USP (Raumdata, 2020, np.).

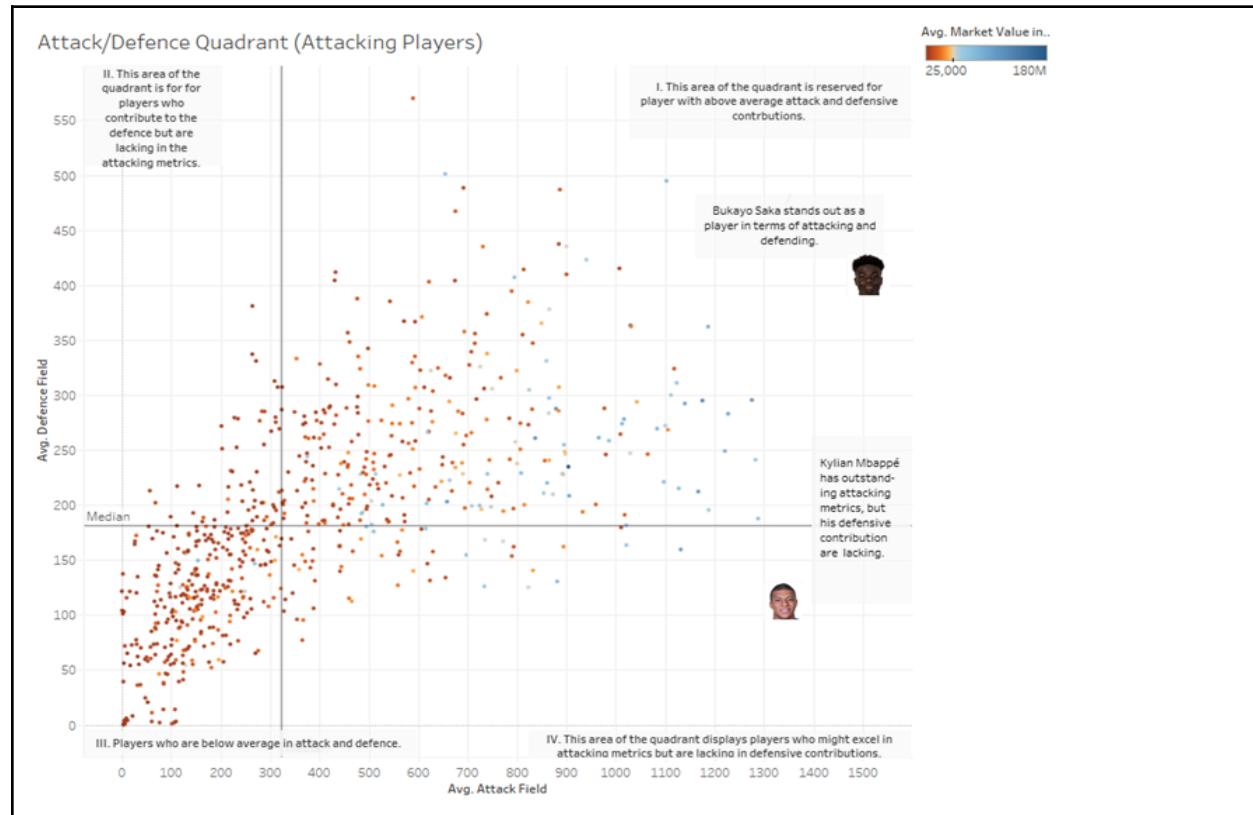
Next, in football, one could suggest that there is a bias towards players of certain nationalities (Kilvington, 2013, pp. 587-594). For example, in the UK, players of Asian nationality are often overlooked by scouts in professional clubs as they are viewed as “physically weak” and more suited to non-contact sports such as cricket (Kilvington, 2013, pp. 587-594). Therefore, it would be interesting if highly regarded footballing nationalities such as Brazil are producing as many talented players as their reputation suggests, or if there are other nations that are producing a large amount of talent but have gone unnoticed (Price, 2023, np.) (Poli et al., 2017, np.).

## Appendix G. Experiments, Evaluation, Results and Analysis

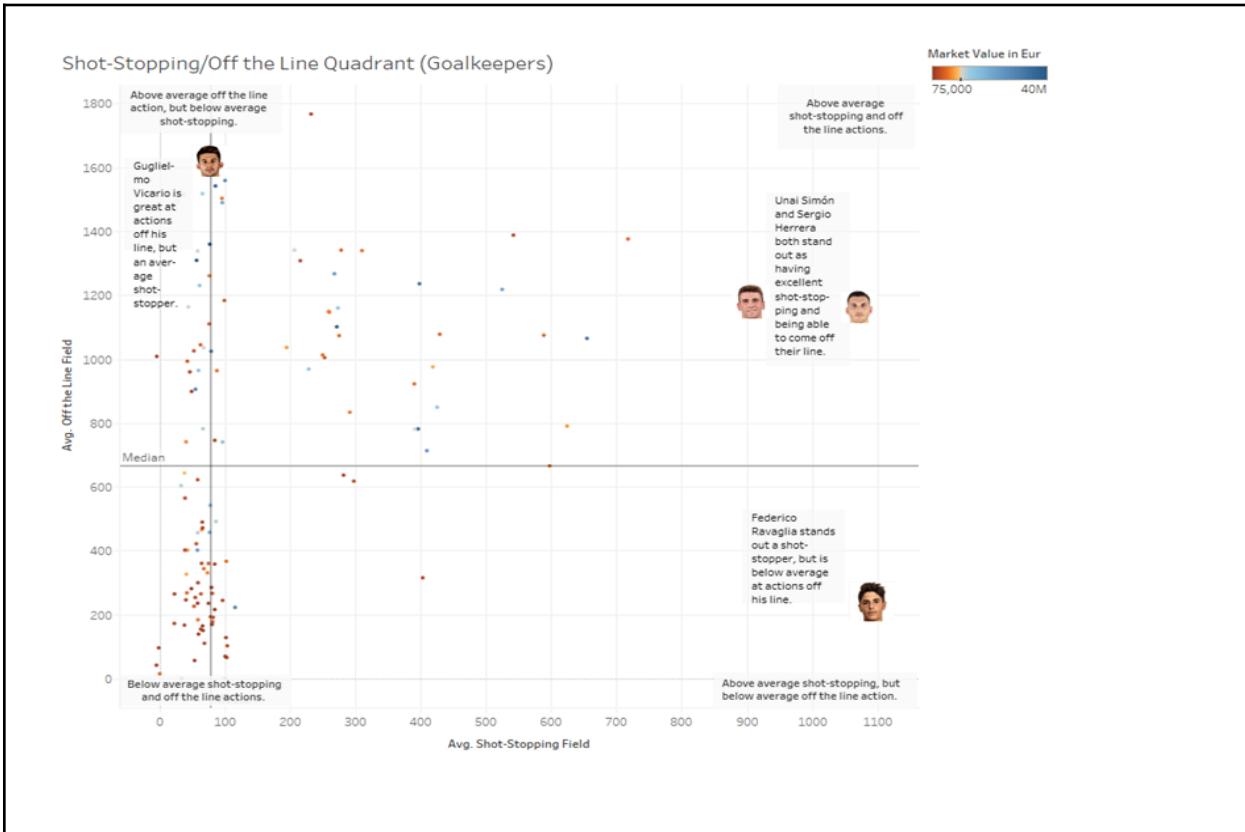
### Appendix G. Application 3: Attack/Defence 4 quadrant grid

The attack defence quadrant graphs comprise of an “attack” and “defence” field which are composites statistics or “aggregate metrics” made up of variables that relate to “attacking” or “defending” (see variable table in appendix C for detailed equations). These are then plotted on graphs with the quadrant being split by the median, separating those who are average, above average and below average in each aggregate metric. For example, with those above both medians excelling in both attack and defence.

### **Figures G.1-G.4**







The Attack/Defence Quadrant managerial application provides insight into the standout qualities and drawbacks of each player. While the PS metric developed by the author weighs both attacking and defensive metrics, with different metrics varying for different positions, the Attack/Defence Quadrant utilises two composite statistics: the Attacking Field and Defensive Field. For goalkeepers, due to their specialised metrics, a Shot-Stopping Field (how good they were at making saves, saving penalties, etc.) and an Off-The-Line Field (how good they were at catching crosses, their passing accuracy, etc.) were used.

These two fields use metrics relating to only measuring their specific quality; for example, the Attacking Field uses only attacking metrics (Goals, Assists, penalties won, etc.) to get a purely offensive advanced metric. Using these two metrics, the author uploaded the data into Tableau (nd., np.) for each position, plotted the players on a scatter plot, and added lines of reference using the median to create a quadrant graph.

The quadrant for outfield players was split into 4 areas: 1) above-average attacking and defending; 2) above-average attacking; and 3) below-average defending; 3) above-average defending, below-average attacking, and 4) below-average attacking and defending.

The quadrant for goalkeepers was also split into four areas: 1) above-average shot-stopping and off-the-line actions 2) above-average shot-stopping and below-average off-the-line actions 3)

below-average shot-stopping and above-average off-the-line actions; and 4) below-average shot-stopping and off-the-line actions.

For attacking players, a player that stood out as having excellent attack and defensive qualities was Bukayo Saka, a winger who has a high tally of goal contributions but has a proven work rate and has proven his willingness to run back after his team loses the ball in order to help his fullback, which helps contribute to the defence. Another standout player was Kylian Mbappé, who excelled offensively with his G+A and ability to dribble past players; however, he lacked defensively due to PSG over the years building their system around his ability to contribute offensively but also make up for his lack of defensive contributions (Tharme, 2024, np.).

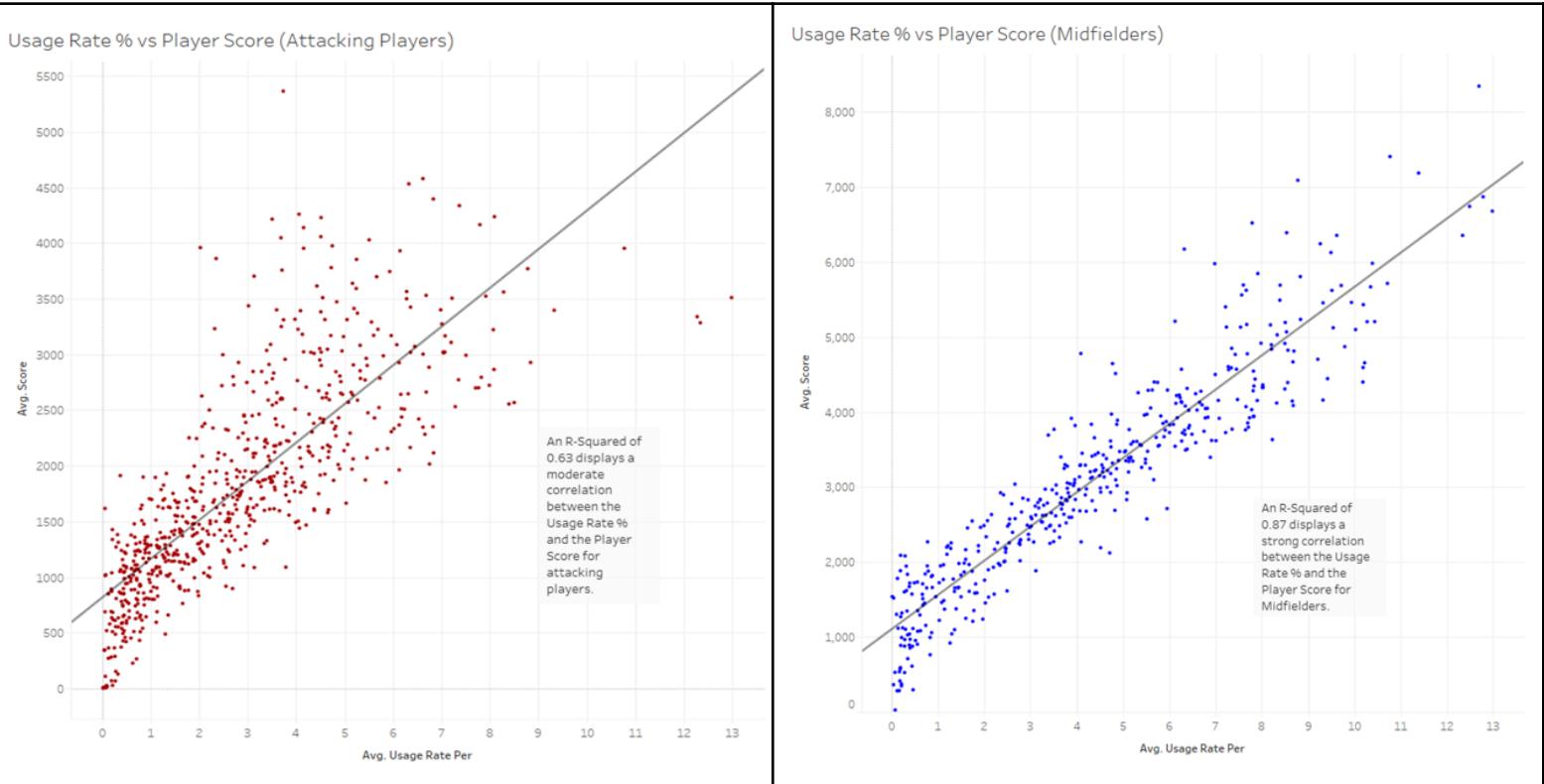
The standout midfielders, both attacking and defensively, were Rodri and Bruno Fernandes. Despite Rodri being classified as a “defensive” and Fernandes being an “attacking” midfielder, they both excel in a midfielder’s traditional role of being excellent in both their defensive end of the pitch and the attacking side. They both progress the ball through carries and passes, score goals, put in tackles, and win the ball through interceptions at a high rate (FBref, nd., np.). Cole Palmer excels in attacking and acts more like a traditional “number 10” or attacking midfielder by having a high rate of goal contributions and doing well at progressing the ball; however, he does lack in the defensive side, potentially due to Chelsea’s attack almost solely relying on Palmer; therefore, the tactics will rely on other players to make up for his lack of defending (Spencer, 2024, np.) (Kinsella, 2024, np.). João Palhinha, on the other hand, acts more like a traditional defensive midfielder, as he excels on the defensive side by making a high number of tackles and interceptions, making ball recoveries, winning duels in the air, etc., but is average attacking-wise.

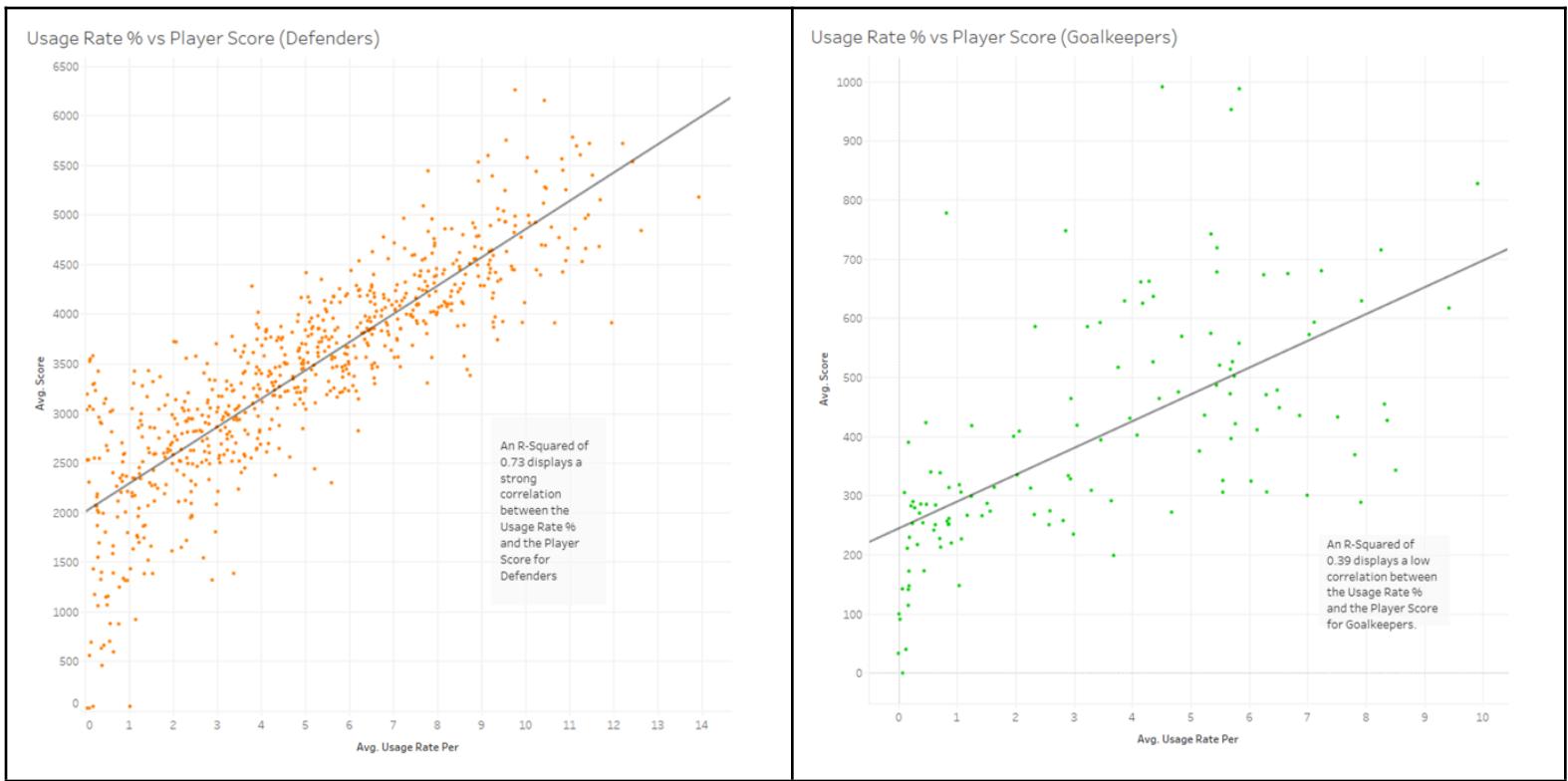
For defenders, a player that stood out as having excellent attack and defensive qualities was Pedro Porro, who under Tottenham’s system acts as an “inverted fullback,” which is when a fullback “inverts” into the midfield to help contribute offensively, e.g., to progress the ball; additionally, he is also a natural “wing back,” therefore progressing the ball upwards towards the opponent goal comes naturally to him, allowing him to excel both offensively and defensively (Tweedale et al., 2023, np.). Federico Dimarco is a player who is above average attacking wise but below average defending wise according to the graph, potentially due to his role as a “wing back,” which gives him more attacking freedom than a traditional fullback, as well as Inter Milan playing with three centre backs rather than one, giving Dimarco defensive cover to allow him to go further up the pitch to attack and progress the ball (The Football Analyst, 2024, np.). On the other hand, Berat Djimsiti is a great centre back, especially one-on-one, but lacks attacking ability, which could be because his team, Atalanta, plays with three centre backs and two wings backs, leaving the centre backs to defend more and allowing the wing backs to go forward (The Football Analyst, 2024, np.).

Finally, for goalkeepers, the standout players in both shot-stopping and off-line actions were Unai Simon and Sergio Herrea, who both excel at making saves but are also able to come off their line and contribute to open play and on-the-ball actions, e.g, they can help their team build up by coming out and effectively working as another defender, along with coming off their line to take on opposition players or to catch crosses from the opposition in a role known as a “sweeper keeper” (Ghar et al., 2021, p.1163). A primary example of a sweeper keeper is Vicario, who is excellent at coming off his line but is lacking in his shot-stopping ability, as displayed by the graph he is on, where he lies on the average line for shot-stopping (Ghar et al., 2021, p.1163). A goalkeeper who excels at making saves but is underwhelming at coming off their line is referred to as a shot-stopper, and on the graphs, the primary shot stopper is Federico Ravaglia, who excels in this area but is below average off the line (Ghar et al., 2021, p.1163).

## Appendix G. Application 4: ‘Usage Rate Percentage’ and its’ impact on a player’s performance.

**Figures G.5-G.9**





The next managerial application is using linear regression to determine whether the ‘USP’ metric has an impact on players’ performances (Raumdata, 2020, np.). USP is a basketball metric that has been adapted to football by Raumdata (2020, np.) using the formula: (Individual player’s touches / Total team touches) \* 100. A touch in this case is defined as a player “receiving a pass, dribbling, and then sending a pass or shot” (Raumdata, 2020, np.).

One could argue that players who have the ball more are generally more effective, as they have more opportunities to score goals, carry or pass the ball progressively, etc. (Raumdata, 2020, np.). Using the processed data, the author uploaded it to Tableau (nd., np.) and created four linear regression graphs for each position to observe if there was a correlation between the players’ USP and their performance, which in this case is measured by their ‘Player Score’ (Raumdata, 2020, np.). The table below and the graphs above display the following results:

**Table G.1**

Position	R-Squared	Correlation
Attacking Player	0.63	Moderate
Midfielder	0.87	High
Defender	0.73	High

Goalkeeper	0.39	Low
------------	------	-----

From the result, one can observe that midfielders have a higher USP, which correlates well with better performance (Raumdata, 2020, np.). This could be because midfielders are very important for building up play and on-the-ball actions such as progressive carries, as midfielders are generally tasked with progressing the ball up the pitch, therefore they are on the ball a lot (giving them a high USP), and the metrics that they excel in, such as progressive passes and carries, improve their PS, leading to in theory a strong correlation (Mota et al., 2023, pp.14-16) (Raumdata, 2020, np.).

The next highest correlation is with defenders, and despite many of their most important metrics being off-ball actions (such as tackles and interceptions), there is still a strong correlation between their USP and their PS (Mota et al., 2023, pp.12-16) (Raumdata, 2020, np.). This could be because over the last decade, defenders have become much more important for their on-ball ability, such as progressing the ball up the pitch, due to the rise of the strategy of “playing out from the back,” which is when teams begin passing and progressing the ball from deep in their own defensive third, requiring defender to be on the ball more and having more touches, increasing their USP, but also placing more emphasis on their on-ball metrics, such as progressive carries (Robson, 2019, np.) (Mota et al., 2023, pp.12-16) (Raumdata, 2020, np.). Therefore, while their off-the-ball metrics are still important, defenders and their USP and PS have become more correlated due to their increasing function in on-the-ball actions (Robson, 2019, np.) (Raumdata, 2020, np.).

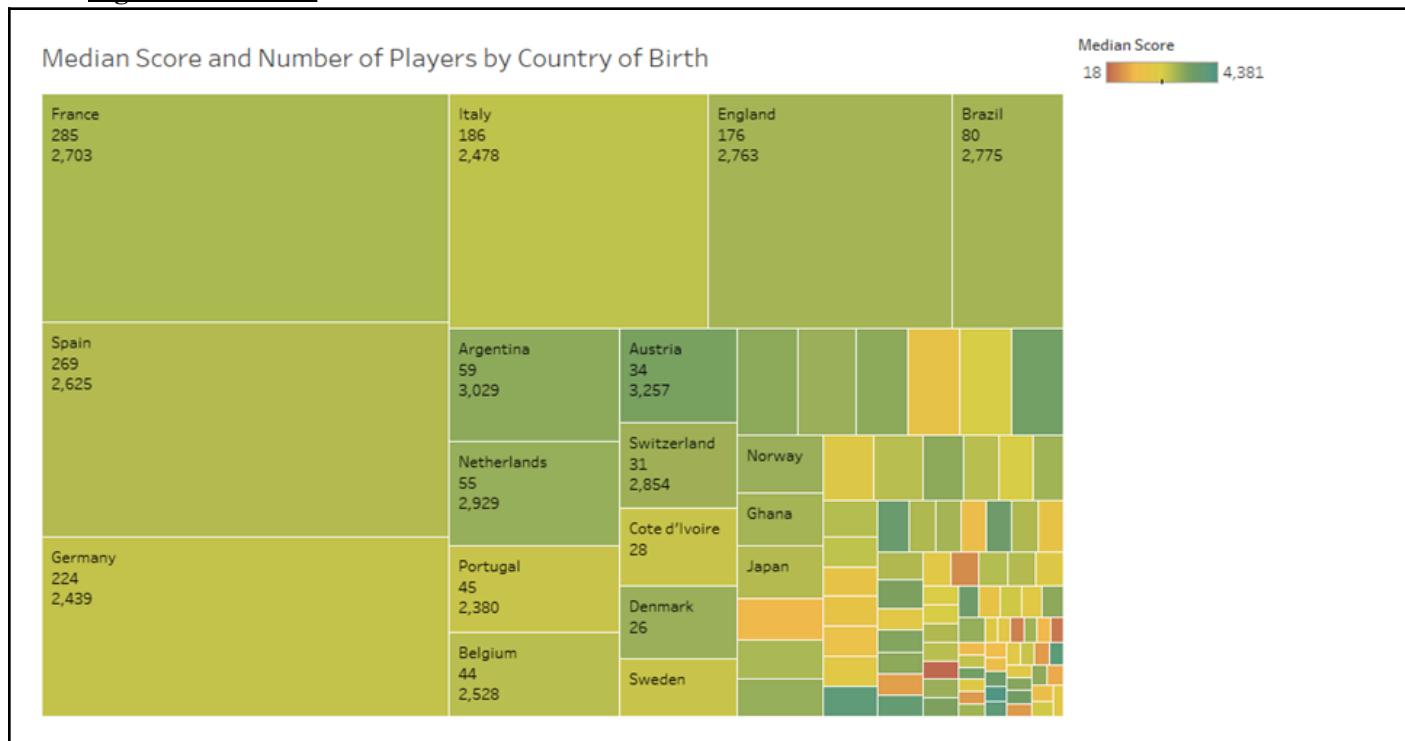
Next is attacking players, where there is a moderate correlation as attacking players (mainly wingers and attacking midfielders) contribute to build-up play such as progressing the ball up the pitch towards the goal, increasing their USP (Raumdata, 2020, np.) (Mota et al., 2023, pp.12-16). However, one could argue that since strikers are generally tasked with playing further on the pitch, they are less active in the build-up play, leading to fewer touches on the ball, with the few touches they do get leading to shots instead of passing back and forth between team members (Mota et al., 2023, pp.6-9, pp.14-16) (Raumdata, 2020, np.).

Finally, goalkeepers have a low correlation between their USP and their PS (Raumdata, 2020, np.). This could be because a goalkeepers' most valuable actions, such as making saves, are off-ball actions, so their quality has very little to do with the number of touches they have on the ball (Raumdata, 2020, np.). Despite the rise of goalkeepers who have good on-ball ability and the increasing use of goalkeepers to help build up from the back, their off-ball ability is still the most reflective of their playing ability and performance (Robson, 2019, np.) (Mota et al., 2023, pp.6-9, pp.12-14).

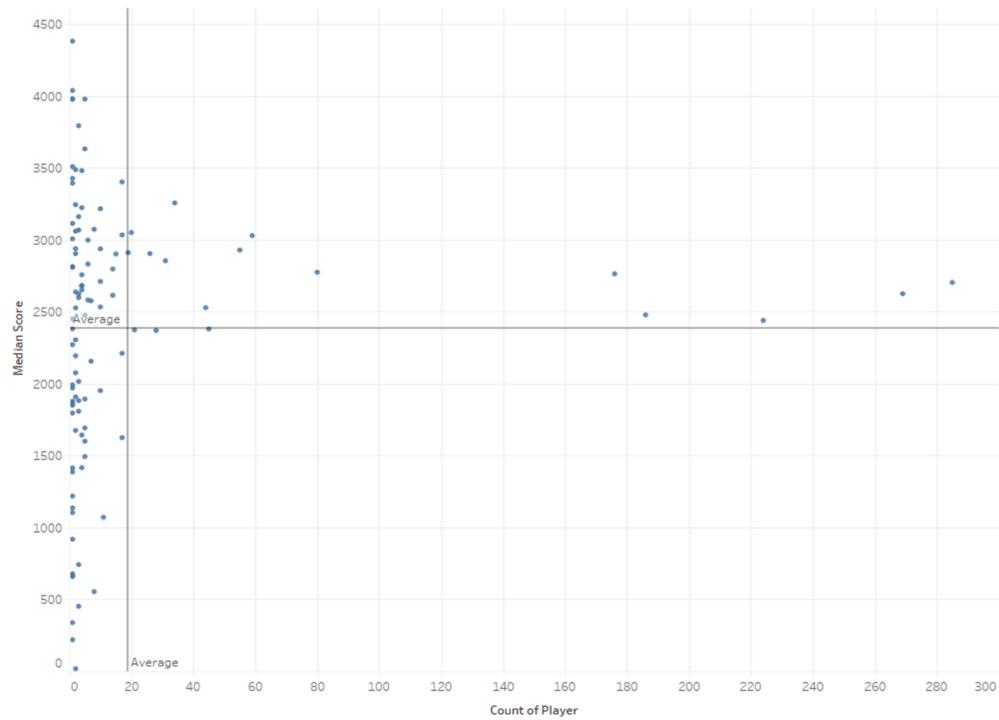
Therefore, from the author's results and analysis, it could be suggested that for certain positions, USP does seem to have a strong correlation with a player's overall performance, especially for players like midfielders who rely on having the ball a lot to win games (Mota et al., 2023, pp.12-16).

#### **Appendix G. Application 5 Create a visualisation map of the nationalities with the most and/or the best players. Countries with the best talent ID.**

## **Figures G.10-G.13**

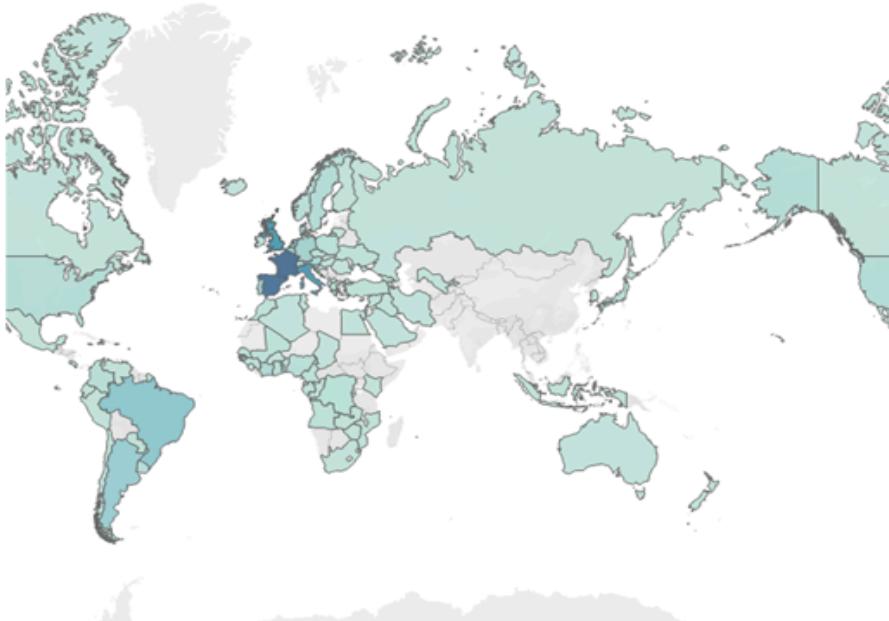


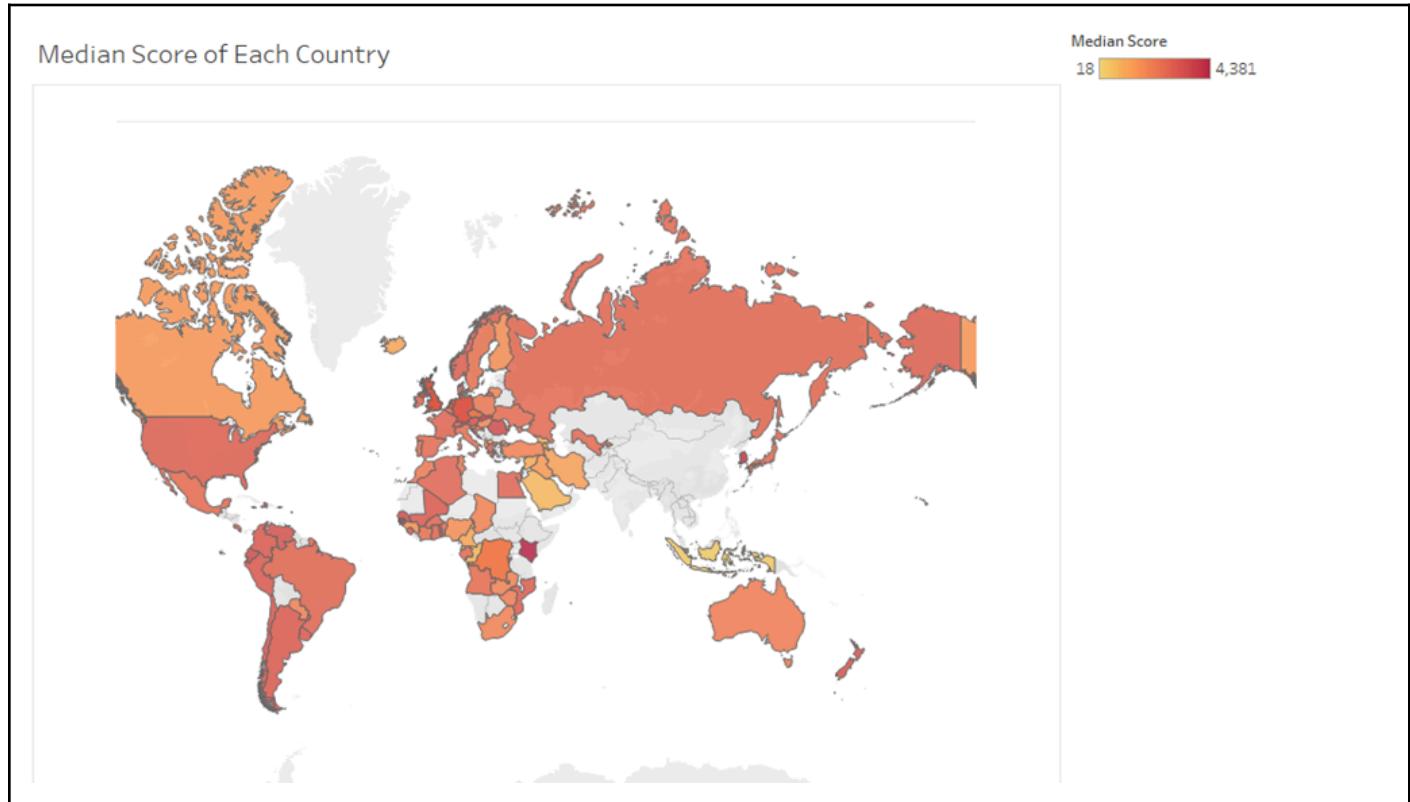
Median Score and Number of Players by Country of Birth



Number of Players in Europe's Top 5 Leagues

Count of Player  
1 285





Using Python (nd., np.), the author took the completed datasets of the 4 positions and joined them together with variables: player, PS and country of birth. Using these variables and Tableau (nd., np.), the author aimed to find the countries that produce the most and/or best players. The graphs visualise the countries with the most players in Europe's Top 5 leagues and the median score for each country, with the median being used as the average to avoid certain countries having the average PS being skewed by outliers (FBref, nd., np.).

From the visualisation above, one can observe that the 5 countries with an above-average number of players and PS are: France, Spain, Germany, Italy, and England. These countries are traditionally considered five of the biggest footballing nations; all the countries have won at least one World Cup and European Championship (except England, which just has one world cup); all the countries are in the top 15 of FIFA's world rankings; and the domestic leagues of the countries are part of the UEFA top 5 European leagues (FIFA, 2024, np.) (FBref, nd., np.) (UEFA, nd., np.). Therefore, with their rich history and relatively high wealth, it makes sense that these countries are above average in their capability of producing talent (FIFA, 2024, np.) (FBref, nd., np.) (UEFA, nd., np.). As these countries' domestic leagues make up Europe's top 5 leagues, it makes sense that they produce the most talent for the top 5 leagues (which is the data from the FBref dataset), as these countries will naturally have more players of their nationality due to the local scouting and academies (FBref, nd., np.). For example, Spanish teams will scout and have academies full of Spanish players who will then go on to become footballers in La Liga

due to them not needing a transfer fee as they are academy graduates, and they will likely stay in the country as transferring to a different league in a different country could lead to language and/or cultural issues (UEFA, nd., np.) (Player Scout, nd., np.).

Brazil, Argentina, and the Netherlands are also above average in these two categories because, like the previous five, they also have a rich football history and heritage (FIFA, 2024, np.) (Taylor, 2006, pp.16-19, pp.28-29. However, the level and number of talent produced is at a lower level than the other five nations, in part because since these three nations also have popular and historically significant football leagues, a large amount of their talent is thriving in their domestic leagues, e.g, there will be more Spanish players than Argentine players in La Liga (Spain's domestic league) due to academy players and local scouting, and vice versa (Taylor, 2006, pp.28-29).

Countries that rank low in the number of players and median PS are Asian countries such as Indonesia, Syria, etc., and smaller European countries such as Iceland and Czechia. This could be because these countries aren't particularly wealthy, have small populations, do not rank highly in FIFA's world rankings, and do not have a strong football heritage (FIFA, 2024, np.) (UEFA, nd., np.).

Some countries with an above-average player but a low number of players are Senegal, Scotland, and Japan. The author chose these three countries as there are different reasons for each country having high-quality players, but very few in Europe's top 5 leagues. Senegal (and many other African countries) have high-quality players but a low number in Europe's top 5 leagues (FBref, nd., np.). This could be because of issues such as the fact that these countries have below-average GDP per capita, resulting in many issues that lead to the talent pool being overlooked, such as potential superstar players being unable to make it because of malnutrition, insufficient infrastructure, etc., (Ungruhe and Schmidt, 2020, p.413-416). In Scotland's case, it could be because of two reasons: it has a small population, and it already has a thriving football league, resulting in talented Scottish players generally staying local but going to one of the two biggest teams (either Celtic or Rangers) (UEFA, nd., np.) (Campbell et al., 2022, np.). In Japan's case, it could be because of the issues mentioned earlier by the author that scouts in England (and presumably the other top 5 leagues) players of Asian nationality are overlooked because of stereotypes, such as South Asian players being viewed as less physical, not being suited to contact sports, and should play a sport like cricket instead (Kilvington, 2013, pp. 587-594).

## **Appendix G. Discussion**

From a managerial perspective, the Attack/Defence graphs provide useful information, such as the areas where certain players excel and could fit on a certain team. For example, a manager with a budget of €30 million might want a player under the age of 30 who excels in attack and defence, so using the graphs, they could observe that a player that fits this criteria is Anthony Elanga, who is 22. His market value is €22 million, and his Attacking and Defensive fields are both above average.

The Attack and Defence fields are also useful composite metrics for a general understanding of a player's attacking and defensive capabilities. This managerial application remains in line with the Moneyball principles set out in the project, as this is a low-cost and effective data-driven means to scout players by using composite statistics with undervalued statistics to identify the players a manager needs (Lewis, 2003).

For scouts and managers, the USP could be useful for scouting players as it demonstrates their ability to be involved in the game, as the linear regression shows, especially for midfielders, that a player with a high USP will also perform well (Raumdata, 2020, np.). This is interesting since in the original PS function, USP was not even included (Raumdata, 2020, np.).

Interestingly enough, in the ranking of players' features for their Predicted Performance, USP ranked second for defenders and fourth for midfielders for the algorithm predicting their performance (Raumdata, 2020, np.).

This could be for a variety of reasons, including that defenders and midfielders are important in the build up play, so having a high number of touches is important for a team to perform well and therefore for them to have good performances (Mota et al., 2023, pp.14-16) (Robson, 2019, np.). Additionally, while a striker does not need to have the ball a lot of to do well (even though for attacking players it ranked fourteenth), for a midfielder and defender, when they have the ball a lot, it results in progressive passes and carries, which leads to a high PS and improved Predicted Performance, as defender who do not have the ball a lot because they either give it away or get subbed off because they are underperforming, leading to worse PS and Predicted Performance (Mota et al., 2023, pp.14-16) (Robson, 2019, np.).

Therefore, USP is a strong indicator of performance, as demonstrated by both linear regression and the random forest algorithm (Raumdata, 2020, np.) (Beheshti, 2022, np.).

The Moneyball country map could be useful for managerial applications, as it demonstrates to managers which countries are already producing a high level of talent (such as France), but also which countries produce very talented players but are being overlooked and undervalued and that more scouts should be sent there (Lewis, 2003). For example, Senegal has produced some very high-level players, but because of socio-economic issues, VISA issues or a lack of eyes on the

country, a large amount of football talent is overlooked (Ungruhe and Schmidt, 2020, pp.413-416). Because of this, a manager could scout Senegalese players or send scouts to Senegal and potentially acquire some talented players for lower transfer fees than an English player would go for.

## **Appendix H. Use of Generative AI statement**

Furthermore, I acknowledge the use of OpenAI (source: <https://chat.openai.com/>) to aid me in making sure my code was efficient and working, using it to help me with errors, structure and understanding the code, which was included in the work submitted for assessment. I confirm that all use of AI content is acknowledged and referenced appropriately.

1)The following prompt was input into (OpenAI:

```
[“# creating a list with new names
new_columns = []
for col in df.columns:
if 'level_0' in col:
    new_col = col.split()[-1] # takes the last name
else:
    new_col = col
new_columns.append(new_col)

# rename columns
df.columns = new_columns
df = df.fillna(0)
```

df.head()” This code is not generating an output with the desired dataset, what is the error? And how could I fix it?]

The output obtained was:

(“The issue seems to be in your loop where you're creating the list new\_col for renaming columns. It appears that the loop is intended to rename columns by removing 'level\_0' from column names if present. However, it seems there's a typo in the variable assignment inside the loop (new\_col\_name vs new\_column\_name).

Here's the corrected code:

```
python
Copy code
# creating a list with new names
new_columns = []
for col in df.columns:
if 'level_0' in col:
    new_col = col.split()[-1]
    new_columns.append(new_col_name) # Fixed variable name here
else:
    new_columns.append(col) # Append unchanged column name if 'level_0' is not present

# rename columns
df.columns = new_col
df = df.fillna(0)
```

```
df.head()
```

With this correction, the code should work as intended, renaming columns by removing 'level\_0' if present and filling any NaN values with 0.”)

Using this corrected output I adapted it to my code including by renaming the variable, using my own web scrapped datasets as the input, etc.

2)The following prompt was input into (OpenAI:

[“For a scouting system, what kind of IPY widget would you recommend? One where I can interact with the scouting system and change the input to give me a different output?”]

The output obtained was:

(“For a scouting system where you want to interact with the input and dynamically see the output, an interactive IPyWidgets dashboard would be ideal. Here's a recommended setup:

#### 1. Dropdown>Select Widget

- Use case: Select options like player position, team, or skill level.
- Widget: Dropdown or SelectMultiple

Example:

```
python
import ipywidgets as widgets
position = widgets.Dropdown(
    options=['Goalkeeper', 'Defender', 'Midfielder', 'Forward'],
    value='Forward',
    description='Position:',
)
●
```

#### 2. Slider Widget

- Use case: Adjust parameters like age, height, weight, or skill rating.
- Widget: IntSlider or FloatSlider

Example:

```
python
Copy code
age_slider = widgets.IntSlider(
    value=25,
    min=16,
    max=40,
    step=1,
    description='Age:',
```

)

•

### 3. Text Input Widget

- Use case: Input player name, team name, or other custom search terms.
- Widget: Text

Example:

python

Copy code

```
player_name = widgets.Text(  
    value=","  
    placeholder='Enter player name',  
    description='Player Name:',  
    • )"
```

I adapted this code by selecting the slider widget and after some research built on the code that ChatGPT provided me to build a fully functional scouting system with my own variables and adapted to my dataset.